PREDICTION OF RNA SECONDARY STRUCTURE IN

HEPATITIS C AND RELATED VIRUSES

ANDREW TUPLIN

A thesis submitted for the Degree of Doctor of Philosophy

The University of Edinburgh

2003

To my family

I declare that the studies presented here are the result of my own independent investigation. This work has not been submitted for any other degree.

Andrew K. Tuplin

ACKNOWLEDGMENTS

I would like to thank my supervisor Prof. Peter Simmonds for all his help, advice and encouragement throughout the work carried out in this thesis. I would also like express my gratitude to all my colleges in the Virus Evolution Group.

My special thanks go to Catriona, my parents Mike and Julie and the rest of my family for all their love, guidance, encouragement (and vast amount of patience).

CONTENTS

CHAPTER 5. VISUALISATION OF RNA STRUCTURE

PUBLICATIONS:

Cuceanu,N.M., Tuplin,A., and Simmonds,P. (2001). Evolutionarily conserved RNA secondary structures in coding and non-coding sequences at the 3' end of the hepatitis G virus/GB- virus C genome. J. Gen. Virol. *82*, 713-22.

Tuplin,A., Wood,J., Evans,D.J., Patel,A.H., and Simmonds,P. (2002). Thermodynamic and phylogenetic prediction of RNA secondary structures in the coding region of hepatitis C virus. RNA *8*, 824-841.

PUBLICATIONS SUBMITED:

Simmonds,P., Tuplin,A., and Evans,D.J., (2003). Pathogenic and evolutionary significance of extensive RNA secondary structure in RNA viruses. Nature Genetics. Submitted for review.

Tuplin,A., Evans,D.J., and Simmonds,P., (2003). Nuclease mapping and covariance analysis of RNA secondary structures in core and NS5B coding region sequences of Hepatitis C virus. J. Gen. Virol. Submitted for review.

ORAL PRESENTATIONS:

Tuplin,A., and Simmonds,P., (2002). Computational and experimental prediction of secondary structure in HGV/GBV-C. Society for General Microbiology, Warwick University. March 2002.

Tuplin,A., and Simmonds,P., (2002). Nuclease mapping and direct visualisation of RNA structure in HCV and HGV/GBV-C. Hepatitis C virus workshop weekend. Keswick. November 2002.

POSTER PRESENTATIONS:

Tuplin,A., and Simmonds,P., (2001). Prediction and visualisation of RNA secondary structure in single stranded RNA viruses. 8th International Symposium on Hepatitis C Virus and Related Viruses. Paris. September 2001.

Tuplin,A., Evans,D.J., Patel,A.H., and Simmonds,P. (2003). Phylogenetic analysis and physical mapping of RNA secondary structure in the core and NS5B regions of HCV. 11th International Symposium on viral hepatitis and liver disease. Sydney. April 2003.

ABSTRACT

The existence and functional importance of RNA secondary structure in the replication of positive-stranded RNA viruses is increasingly recognised. In this thesis several computational methods to detect RNA secondary structure in the coding regions of hepatitis C virus (HCV), hepatitis G virus (HGV)/GB virus C (GBV-C) and related viruses have been used. These include thermodynamic prediction of folding free energies (FFEs), evolutionary conservation of minimum energy structures between virus genotypes, suppression of synonymous variability and analysis of covariant and semi covariant substitutions in thermodynamically favoured structures. Each of the predictive methods provided evidence for conserved RNA secondary structure in the core and NS5B encoding regions of HCV and throughout the entire coding region of HGV/GBV-C.

Positions in the HCV genome with predicted RNA structure localise precisely to regions of marked suppression of variability at synonymous sites, indicating that RNA structure constrains sequence change at what are generally regarded as phenotypically neutral sites. Combining these methods, the computational data obtained in this thesis demonstrates the existence of at least ten conserved stem loop structures within the NS5B coding region and three in that coding for the core protein both within the coding region of HCV. Analysis of the NS5B coding region and 3' untranslated region (3'UTR) of HGV/GBV-C indicates an even greater degree of RNA secondary structure. Remarkably, it appears from analysis of FFEs that extensive RNA secondary structure may exist along the entire length of both the

HCV and HGV/GBV-C genomes, a finding with considerable implications for future functional studies.

The existence of predicted RNA structures in the HCV genome was determined using controlled nuclease mapping of RNA transcripts from the core and NS5B regions under conditions which retained potential long-range RNA interactions. The pattern of cleavage sites of nucleases specific for single and double stranded RNA provided strong experimental support for structures previously predicted in this study. Electron microscopy was also used to directly visualise the RNA folding structure of HGV/GBV-C and provided some evidence for at least four structures within the NS5B coding region and long range RNA folding across the length of the virus genome.

The degree of structural conservation between diverse HCV and HGV/GBV-C genotypes and related viruses suggests roles in virus replication, and/or RNA packaging for the discrete structures identified in this thesis. Whilst this role and that of the genome wide structure identified is currently not understood the structures predicted in this work are providing a starting point for such functional studies using the HCV replicon.

ABBREVIATIONS

Alanine aminotransferase	ALT
Antisense primer	AS
Base pair	bp
Bovine viral diarrhea virus	BVDV
Codon order randomisation	COR
Complementary DNA	cDNA
Cytomegalovirus	CMV
Degrees Celsius	°C
Deoxyribonucleic acid	DNA
Deoxyribonucleotides	dNTPs
Dinucleotide randomisation	CDR
Dinucleotide swap	CDS
Enzyme linked immunosorbent assay	ELISA
Escherichia coli	*E. coli*
Endoplasmic reticulum	ER
Epstein-Barr virus	EBV
Folding free energy	FFE
Folding free energy difference	FFED

GB virus A	GBV-A
GB virus B	GBV-B
GB virus C	GBV-C
Hepatitis A virus	HAV
Hepatitis B virus	HBV
Hepatitis C virus	HCV
Hepatitis D virus	HDV
Hepatitis E virus	HEV
Hepatitis F Virus	HFV
Hepatitis G Virus	HGV
Hepatocellular carcinoma	HCC
Hypervariable region	HVR
Internal ribosome entry site	IRES
Interferon	IFN
Intravenous drug user	IVDU
Like codon randomisation	CLR
Like codon swap	CLS
Microgram	µg
Microlitre	µl
Milligram	mg

Millilitre	ml
Millimolar	mM
Multiple cloning site	MCS
Nanometre	nm
Non-A, non-B hepatitis	NANBH
nucleotides	nt
Nucleotide order randomisation	NOR
Outer primer	O
Open reading frame	ORF
Polymerase chain reaction	PCR
Reverse transcription polymerase chain Reaction	RT-PCR
Revolutions per minute	rpm
Ribonucleic acid	RNA
Reverse transcription	RT
RNA dependent RNA polymerase	RdRp
Sense primer	s
Saccharomyses cerevisiae	*S. cerevisiae*
Transfusion associated hepatitis	TAH
Untranslated region	UTR

Ultra violet	UV
Yellow fever virus	YFV
5' Untranslated region	5'UTR
3' Untranslated region	3'UTR

CHAPTER 1

INTRODUCTION

1 INTRODUCTION

1.1 DISCOVERY OF *HEPATITIS* C VIRUS

Viral hepatitis is a major public health issue throughout the world. Initially, the two major aetiological viral agents of transfusion-associated hepatitis (TAH) were believed to be hepatitis A virus (HAV) and hepatitis B virus (HBV). HAV is single stranded RNA virus which was originally classified as a member of the enterovirus genera (enterovirus 72) and is a major cause of hepatitis epidemics via the faecal-oral route (Melnick 1982). HAV has since been reclassified into the hepatovirus genus within the family *Picornaviridae*. HBV is a double stranded, circular DNA virus, with a circular genome, and is classified in the *Hepadnaviridae* family. In the 1970s serological tests were developed for both viruses, such as anti-core protein antibody or surface antigen (HBsAg) tests for HBV and immunoglobulin M antibody to HAV. After screening of patients it became evident that many cases of hepatitis associated with blood transfusions were not due to either HAV or HBV and were thus described as non-A, non-B hepatitis (NANBH). The acute phase of NANBH is a mild disease resulting in jaundice in less than half of cases. In at least 50% of cases it leads to a chronic infection (compared with 5% in HBV infection) which may result in progression to cirrhosis of the liver and hepatocellular carcinoma (HCC) after a long incubation period.

Both cytomegalovirus (CMV) and Epstein-Barr virus (EBV) are known to cause liver damage and can be transmitted by blood transfusion. However, in a number of

studies both these viruses were discounted by serological testing of TAH patients who were also shown to be negative for both HAV and HBV (Feinstone *at al.*. 1975). Consequently a diagnosis of NANBH was made by excluding HAV, HBV, herpesviruses such as CMV and EBV and other infectious causes of hepatitis. NANBH was documented as representing over 90% of TAH cases in the United States and up to 10% of transfusions were estimated to result in NANBH (Alter *at al.* 1975; Aach *at al.* 1981).

Progress towards identifying the cause of NANBH was hampered by the difficulty in culturing the agent in either organ or cell culture. However, evidence of a viral aetiology was provided in epidemiological studies which suggested that NANBH hepatitis was transmissible between humans (Prince *at al.* 1974; Alter *at al.* 1978). After successful transmission to chimpanzees (Alter *at al.* 1978), experimental evidence from filtration and chloroform inactivation studies was obtained suggesting that the infectious agent was an enveloped particle between 30 and 60 nm in diameter (Bradley *at al.* 1983; Feinstone *at al.* 1983; Bradley *at al.* 1985).

DNA recombinant technology was finally used to create a cDNA clone based on the genome of what would subsequently be known as hepatitis C virus (HCV) (Choo *at al.* 1989). Nucleic acid (both DNA and RNA) was extracted from experimentally infected chimpanzees following ultracentrifugation of large volumes of relatively high infectious titre plasma. The total nucleic acid was then reverse transcribed using random primers and the resulting cDNA fragments cloned in to bacteriophage λgt11 (cDNA library). Recombinant proteins were then expressed as fusion proteins in *Escherichia coli* (*E. coli*) colonies during bacteriophage replication. Sera from patients with chronic NANBH was then used to screen large numbers of colonies;

leading to the eventual identification of cDNA clone 5-1-1. Hybridisation studies showed that the clone was not derived from the genomes of either human or chimpanzee hosts. Further, total nucleic acid derived from NANBH chimpanzee plasma hybridised to clone 5-1-1 after treatment with deoxyribonucleases, but not after digestion with ribonucleases, and only one of the cDNA strands was found to hybridise to the RNA extracted from plasma. These results indicated that the virus has a single stranded RNA genome (Choo *at al.* 1989).

The 5-1-1 clone was subsequently used as a hybridisation probe against the original cDNA library to identify three overlapping clones which were used to reconstruct a single open reading frame (ORF) which was expressed in yeast as a fusion protein (C100-3) (Kuo *at al.* 1989). C100-3 was 363 viral amino acids in length and was used as the basis of serological screening studies, in which purified protein was used to coat the wells of microtiter plates, so that circulating antibody in patient blood samples could be assayed. Detection of bound antibody was achieved by challenge with a radioactive second antibody. The recombinant protein was shown to be consistently recognised by sera derived from NANBH patients but not from control individuals or those with hepatitis of different aetiology. It was also shown that approximately 80% of patients with chronic, post-transfusion NANBH in Italy and Japan had antibodies to HCV, compared to 15% in those cases where hepatitis was observed in acute, resolving infections. In addition, 58% of NANBH patients in the United States, with no known source of parenteral exposure to the virus, were shown to be positive for the HCV antibody (Kuo *at al.* 1989). Retrospective studies of 16 patients with chronic transfusion associated NANBH, in which linked donor and recipient samples were available, showed that 88% of the

recipients had received transfusions from donor plasma in which HCV antibody was detectable (Alter *at al.* 1989). From such data, it became clear that HCV was the causative agent of the majority on transfusion associated NANBH.

Comparative analysis of the genomes of several strains of HCV, reconstructed from overlapping cDNA clones, was used to determine the structure and organisation of the virus (Choo *at al.* 1991; Takamizawa *at al.* 1991; Okamoto *at al.* 1991). It was shown that the virus genome is approximately 9500 nucleotides in length with a single ORF which is flanked by both 5 'and 3' untranslated regions (5' and 3' UTRs). The ORF encodes a single polyprotein approximately 3000 amino acids in length which and is co- and post-translationally cleaved by host and viral proteases into individual mature proteins. The polyprotein itself can be divided into upstream and downstream domains. The upstream domain codes for three main structural proteins core (C), E1 and E2 (envelope proteins) and P7. The downstream domain codes for the non structural proteins NS2, NS3, NS4A, NS4B, NS5A and NS5B.

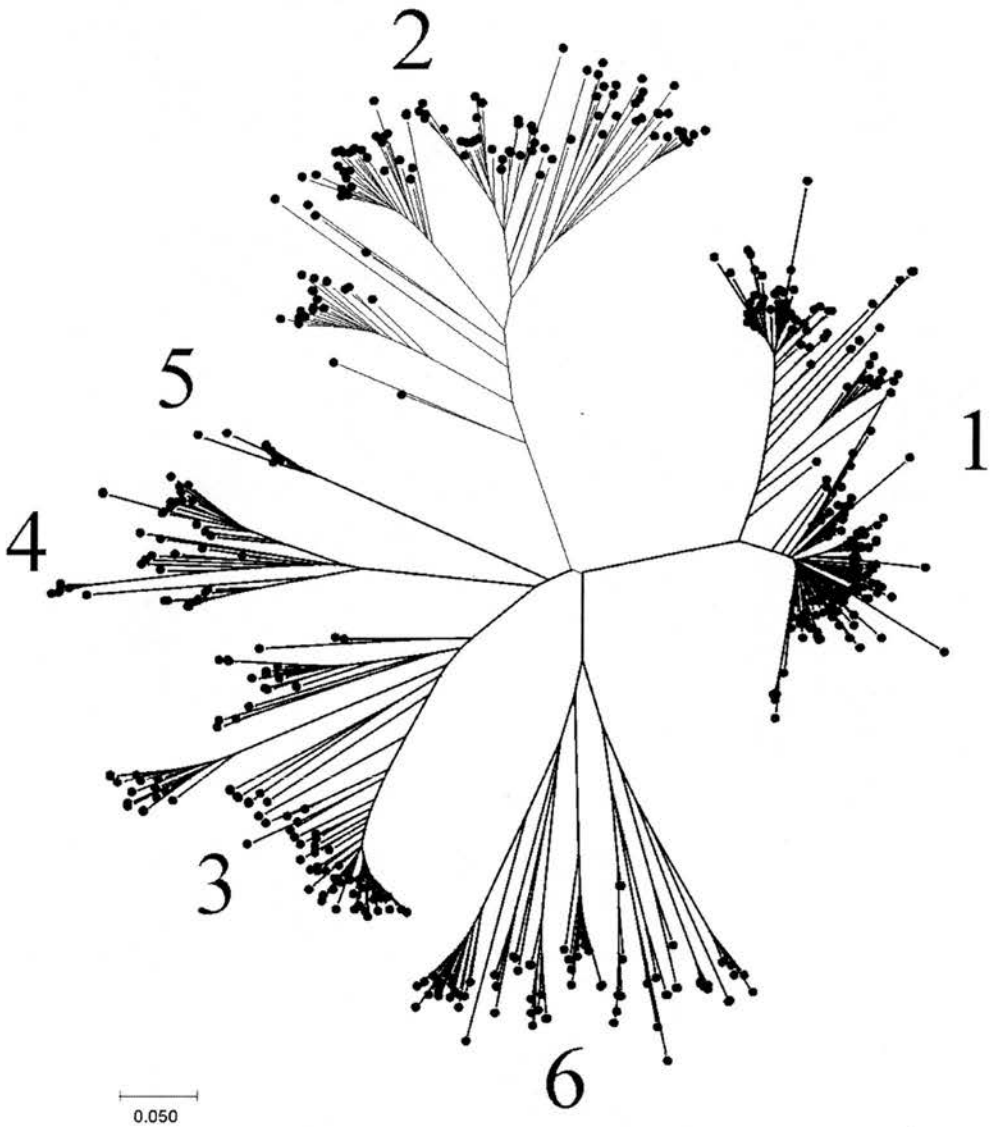
Based on nucleotide and amino acid sequence comparisons and similarities in polyprotein organisation and hydrophobicity profiles, HCV was shown to group phylogenetically with pestiviruses and flaviviruses (Millera and Purcell 1990; Koonin 1991; Choo *at al.* 1991; Takamizawa *at al.* 1991). For example, comparative analysis of the amino acid sequences from single stranded positive sense RNA viruses showed that the RNA dependent RNA polymerase protein (RdRp) (NS5B) contained conserved motifs either side of the active site which grouped into three large subgroups (Koonin 1991). HCV clustered in subgroup II which also included carmoviruses, tombusvirus, luteiviruses, pestiviruses, flaviviruses and unexpectedly, single stranded RNA bacteriophages (Koonin 1991).

It has since been shown by comparative analysis of nucleotide sequences that the genome of HCV is relatively heterogeneous. This has led to the classification of the virus into six major genotypes whose distribution varies both geographically and between risk groups. The genotypes have been numbered 1 to 6 and are equally divergent from each other, differing by 31% to 34% at the nucleotide level and exhibiting approximately 30% amino acid divergence (Fig. 1.1) (Simmonds *at al.* 1994a). The six main genotypes comprise several more closely related subtypes with approximately 20% nucleotide divergence, while within the subtypes variation is less than 10% (Simmonds and Smith 1999a). Different regions of the genome exhibit varying levels of sequence diversity. In particular the non-structural and core coding regions exhibit lower levels of divergence between genotypes than has been observed within the E1 and E2 coding regions. For example, 49% nucleotide sequence divergence has been observed for the E1 protein coding region, between genotypes 1 and 2, as compared to approximately 20% for the core gene (Minutello *at al.* 1993; Bukh *at al.* 1994). However, it has been shown that congruent phylogenetic relationships are observed irrespective of whether the complete genome or discrete regions are analysed (Chamberlain *at al.* 1997).

1.2 SEARCH FOR VIRAL AGENTS OF NON-A,B,C,D HEPATITIS

With the advent of specific serological testing HAV, HBV and HCV molecular cloning techniques led to the isolation of a cDNA clone believed to be a further human hepatitis virus, which was designated hepatitis E virus (HEV) (Reyes *at al.*

Figure 1.1. Phylogenetic analysis of HCV diversity based on analysis of NS5B sequences. Major genotype groupings are labelled in bold type (1-6). Approximately 32% sequence divergence between genotypes is observed with approximately 20 % diversity within genotypes (Reproduced with permission of Peter Simmonds, adapted from Simmonds *et al.* 1999b).



1990). HEV was shown to be a non-enveloped, positive sense, single stranded RNA virus sharing a number of features including sedimentation coefficient and sensitivity to CsCl with caliciviruses (Reyes *at al.* 1990). HEV has since been shown to be a major causative agent of water born epidemics of acute hepatitis.

In 1994 it was reported that novel 27-37 nm particles were observed by electron microscopy in the stool sample of a patient with sporadic NANBH, which when serially inoculated into rhesus monkeys, produced hepatopthic lesions (Deka *at al.* 1994). After inoculation with infectious plasma 27-37 nm particles were subsequently isolated from the infected rhesus monkey livers. The putative infectious agent was reported to possess a double stranded DNA genome, approximately 20 kb in length and was provisionally named hepatitis F virus (HFV) (Deka *at al.* 1994). Subsequent research has failed to confirm these findings.

With the advent of reliable assays for the detection of HAV, HBV, HCV and HEV it became increasingly evident that the aetiology of a substantial fraction of NANBH cases remained undefined, suggesting the existence of additional causative agents (Alter *at al.* 1989) (Simons *at al.* 1995a). For example, in a five year prospective study of 182 adult patients with NANBH in Greece only 47% were shown to be of a non A-D aetiology (Tassopoulos *at al.* 1992). Of these 29.1% acquired the infection after parenteral exposure and 71.9% were community acquired cases, associated with HEV (Tassopoulos *at al.* 1992) (Tassopoulos *at al.* 1994). Further evidence of the existence of an unknown aetiological agent of TAH was provided in a number of studies which reported levels of non-A, non-E chronic hepatitis between 1.5 and 5% (Aach *at al.* 1991) (Hammel *at al.* 1994; Peters *at al.* 1993).

1.3 DISCOVERY OF GB VIRUSES

In 1967 attempts to isolate HAV led to the first report of a "GB hepatitis agent" following experimental inoculation of tamarins (*Saguinus sp.*) with patient sera (Deinhardt *at al.* 1967). The serum was obtained from a 34 year old surgeon (initials GB) from Chicago in the third day of jaundice with an acute mild form of hepatitis and was used to set up a series of five serial marmoset to marmoset passages. Animals directly inoculated with this serum developed hepatitis, as did animals inoculated with serum from tamarins with acute-phase hepatitis after subsequent passage. Hepatitis was assessed by elevated liver enzyme levels in serum and histological abnormalities in the tamarin livers, following biopsy. Based on cross-challenge and differential neutralisation data it was demonstrated that the GB agent was distinct from HAV. Immune electron microscopy was used to directly visualise the infectious agent, from a high dose serum sample (tamarin passage 11) (Almeida *at al.* 1976). The GB agent was observed to have a particle diameter ranging between 20–22 nm as opposed to HAV which has an average diameter of 27 nm.

Further evidence for the novel nature of the GB agent was presented in 1989 after further passage and cross-challenge experiments (Karayiannis *at al.* 1989). Three tamarins (*Saguinus labiatus*), two of which had previously been infected with and were immune to HAV, were inoculated intravenously with plasma containing the GB agent. All three animals developed acute hepatitis within two weeks post inoculation, based on increased levels of serum alanine aminotransferase (ALT) and biopsies showing liver abnormalities. These results further indicated that the GB agent was unique from HAV, as infection was induced in animals with prior

immunity to HAV. In the same study it was shown that faecal specimens taken from acutely infected animals were non-infectious; indicating that the virus was unlikely to be an enterovirus and was distinct from enterically transmitted NANBH hepatitis virus such as HEV.

In 1995 a system of representational difference analysis (RDA) was used to provide a more detailed molecular characterisation of the GB agent genome (Simons *at al.* 1995b). RDA was developed by Lisitsyn and Wigler in order to amplify unique DNA sequences present in one complex source but absent in a highly related one (Lisitsyn *at al.* 1993). Plasma was taken from two tamarins; pre-inoculation with GB agent sera and post inoculation during the acute phase of hepatitis, as measured by increased levels of serum ALT. Reverse transcription followed by second strand synthesis was performed on total extracted nucleic acid using random hexamer primers. The DNA fragments were digested with *Sau*3Ai and ligated to a compatible primer set (Lisitsyn *at al.* 1993). Using RDA, 76 sequences present in the post inoculation, but not the pre-inoculation samples, unique sequences were identified and cloned.

Following cross hybridisation eleven clones were identified which were shown by Southern blot analysis to be absent in pre-inoculation tamarin, human, *Saccharomyses cerevisiae* (*S. cerevisiae*) and *E. coli* genomic DNA indicating an exogenous source. Of these, seven were selected for further investigation all of which could be detected by RT-PCR, but not PCR, in infectious tamarin plasma; suggesting an RNA genome. Sequence analysis of the seven clones, followed by comparison with the GenBank and Swiss-prot nucleotide and protein databases, revealed that 5 had limited amino acid sequence identity to the non structural

proteins of HCV. Extension of the cloned sequences followed by further sequence analysis demonstrated the presence of two distinct virus RNA molecules within the infected tamarin plasma. Both viruses had limited sequence identity to flaviruses and were designated GB virus B (GBV-B) and GB virus A (GBV-A) (Simons *at al.* 1995b).

The genomes of both GBV-A and GBV-B were reconstructed from overlapping cDNA clones, isolated from infectious tamarin serum and liver samples (Muerhoff *at al.* 1995). GBV-A was observed to have a genome 9493 nucleotides in length, encoding a potential polyprotein of 2972 amino acids. The genome of GBV-B was found to be 9143 nucleotides long, encoding a potential polyprotein of 2864 amino acids. Based on a comparison of hydrophobicity profiles the genomes of the two viruses were shown to be organised in a similar way to those of pestiviruses (BVDV), flaviviruses (YFV) and HCV, with predicted structural genes located towards the 5' of the genome upstream of the predicted non structural protein coding region (Muerhoff *at al.* 1995). Amino acid alignments and subsequent phylogenetic analysis of the predicted RdRp and helicase encoding regions showed that they clustered within supergroup II (Koonin 1991), although they could not be considered the same virus or genotypes of HCV and represent a subgenus within the hepaciviruses of the *Flaviviridae* family.

In order to investigate the seroprevalence of antibodies to HCV, GBV-A and GBV-B and epidemiological links of the GB viruses to non A-E hepatitis, serum samples from a large number of individuals were screened using GBV-A and GBV-B specific ELISAs (Simons *at al.* 1995a). The ELISAs used recombinant proteins coded for by the putative core and non structural genes of GBV-B and the non

structural genes of GBV-A. The screened samples were obtained from three diverse sources; a low risk group of United States volunteer blood donors which had previously tested negative for antibodies to HCV, and HBV surface (HBsAG) and core antigens (anti-HBc); intravenous drug users (IVDUs) with a high prevalence of HCV (99%) and HBV (76%) infection and a group of West African samples from where infection with hepatitis agents is considered relatively common. The United States blood donors exhibited a seroprevalence of 3% to GBV-A and 1.2% to GBV-B. In contrast the IVDUs showed a prevalence of 14% to both viruses and the West African subset 8.4% to GBV-A and 14.6% to GBV-B (Simons *at al.* 1995a).

The immunoreactive sera was tested for the presence of virus RNA with degenerate primers, designed to amplify the putative helicase coding regions (NS3) from HBV-A, GBV-B or HCV genomes. A novel amplification product was generated from the West African sample set whose sequence identity was 59.0%, 53.7%, 47.9% at the nucleotide level and 64.2%, 57.3% and 50.4% at the amino acid level to GBV-A, GBV-C and HCV respectively. BLAST homology comparisons of the nucleotide and amino acid sequences did not detect any significant similarity to any further published sequences. After alignment of the amino acid sequence with those of the *flaviviridae* family it was proposed that the sequence was amplified from a novel virus. Due to the relatively high sequence identity with GBV-A the virus was named GB virus C (GBV-C). No amplification product was detected when the RT step was omitted and RT-PCR failed to detect GBV-C sequences in human, Rhesus monkey, *S. cerevisiae* or *E. coli* DNA samples, confirming that the GBV-C sequence was exogenously derived from an RNA virus. Analysis of the near complete genome of the virus through reconstruction of overlapping clones revealed

a sequence approximately 9125 nucleotides in length with a single ORF coding for a putative polyprotein of 2906 amino acids (Leary *at al.* 1996). Comparative analysis of hydrophobicity plots predicted a similar genome organisation to GBV-A and other members of the *Flaviviridae* family, with the structural proteins located towards the N-terminal of a single polyprotein and the non structural proteins downstream towards the C-terminus (Leary *at al.* 1996).

1.4 DISCOVERY OF HEPATITIS G VIRUS

A novel virus, designated Hepatitis G virus (HGV), was identified and characterised by a different group in a parallel search for unidentified aetiological agents associated with non-A-E hepatitis (Linnen *at al.* 1996). The virus was isolated from a patient with chronic NANBH. Initially the patient was believed to be negative for HCV infection, based on the first generation immunoassay but was later shown to be positive by the second generation immunoassay and RT-PCR for the 5'UTR (Linnen *at al.* 1996).

RNA was extracted from the plasma and reverse transcribed using sequence independent single primer amplification. The RNA products were cloned into a λ gt11 library and patient plasma used to screen for immunoreactive clones. Analysis of the immunoreactive clones revealed HCV-related sequences as well as a number which did not match any sequences in the GenBank database, from which primers were designed. PCR was performed on various sequences such as genomic human DNA, *E. coli* and *S. cerevisiae*. No amplification products were obtained, suggesting

an exogenous source for the sequence contained in the immunoreactive clone. No PCR product was obtained from the patient plasma when the RT step of the reaction was omitted, demonstrating that the exogenous sequence was RNA derived.

The initial exogenous sequence was extended using anchored PCR to produce multiple overlapping cDNA clones whose combined sequence length was 9392 nucleotides, with a single ORF encoding a putative polyprotein of 2873 amino acids (Linnen *at al.* 1996). The genome organisation of HGV resembles that of HCV and the GB viruses, with the structural and non-structural proteins coded by the 5' and 3' regions of the ORF and 5' and 3' UTRs. Comparison of the amino acid sequence with members of the *Flaviviridae* family revealed 43.8% sequence identity to GBV-A, 28.4% to GBV-B and 26.8% to HCV. Higher levels of sequence identity were observed when only the RdRp and helicase domains were analysed, although the same trend was maintained. A comparison of the putative NS3 domain with that of GBV-C showed 85.5% nucleotide sequence identity and 100% amino acid identity. Further analysis of complete genome sequences of HGV and GBV-C revealed 85% nucleotide and 95% amino acid sequence identity (Zuckerman 1996). Consequently it was shown that GBV-C and HGV are isolates of the same virus. The terminology HGV/GBV-C will be used throughout this study.

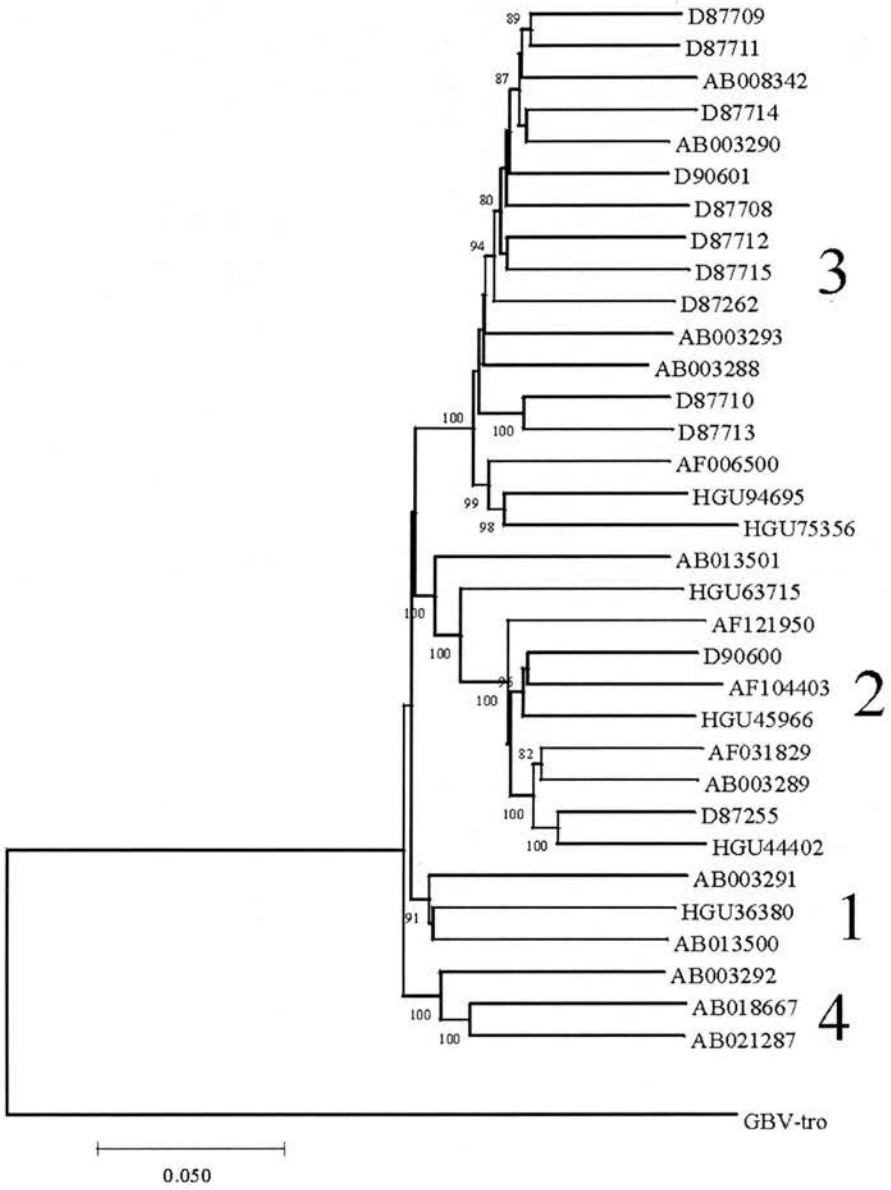
Unlike HCV, variants of HGV/GBV-C show somewhat limited sequence variability. In an analysis of 33 epidemiologically and geographically diverse isolates the most extreme variants of HGV/GBV-C exhibited only 14% sequence divergence across the complete genome sequence (Smith *at al.* 2000). Initial attempts to define genotype groupings relied on analysis of subgenomic fragments such as the NS3 or NS5B encoding regions (Katayama *at al.* 1997; Viazov *at al.*

1997). However, analysis of complete genome sequences failed to reproduce the observed phylogenetic relationships observed between subgenomic fragments (Smith *at al.* 1997a). It was later shown, after analysis of 33 epidemiologically unlinked isolates, that HGV/GBV-C clusters into four major phylogenetic groupings which are geographically distinct and equally divergent from a recently isolated chimpanzee isolate GBV-C_{CPZ} (Fig. 1.2) (Smith *at al.* 2000). These four groupings are not reproducible by analysis of individual genes or subgenomic fragments, with the exception of the E2 coding region (Smith *at al.* 2000).

1.5 HCV GENOMIC ORGANISATION

The genome of HCV is a single stranded positive sense RNA molecule of approximately 9600 nucleotides. The genome carries a single long ORF flanked by both 5' and 3'UTRs and encodes a single polyprotein that is both co- and post-translationally cleaved by viral and cellular proteases (Fig. 1.4). Translation of the polyprotein is directed in a cap independent manner via an internal ribosome entry site (IRES) in the 5'UTR, which acts as a landing pad for the 40S ribosomal subunit, permitting direct binding of the translation initiation complex in close proximity to the start codon of the ORF (Tsukiyama Kohara *at al.* 1992; Wang *at al.* 1993).

Figure 1.2. Phylogenetic analysis of HGV/GBV-C diversity based on analysis of complete genome sequences, rooted with HGV/GBV-C_{CPZ} (GBV-tro). Genotype groupings are labelled in bold type (1-4). Divergence between genotypes is far less than that observed for HCV (approximately 32% for HCV compared to greatest divergence of 13% for HGV/GBV-C) (Reproduced with permission of Peter Simmonds, adapted from Simmonds *et al.* 1999b).

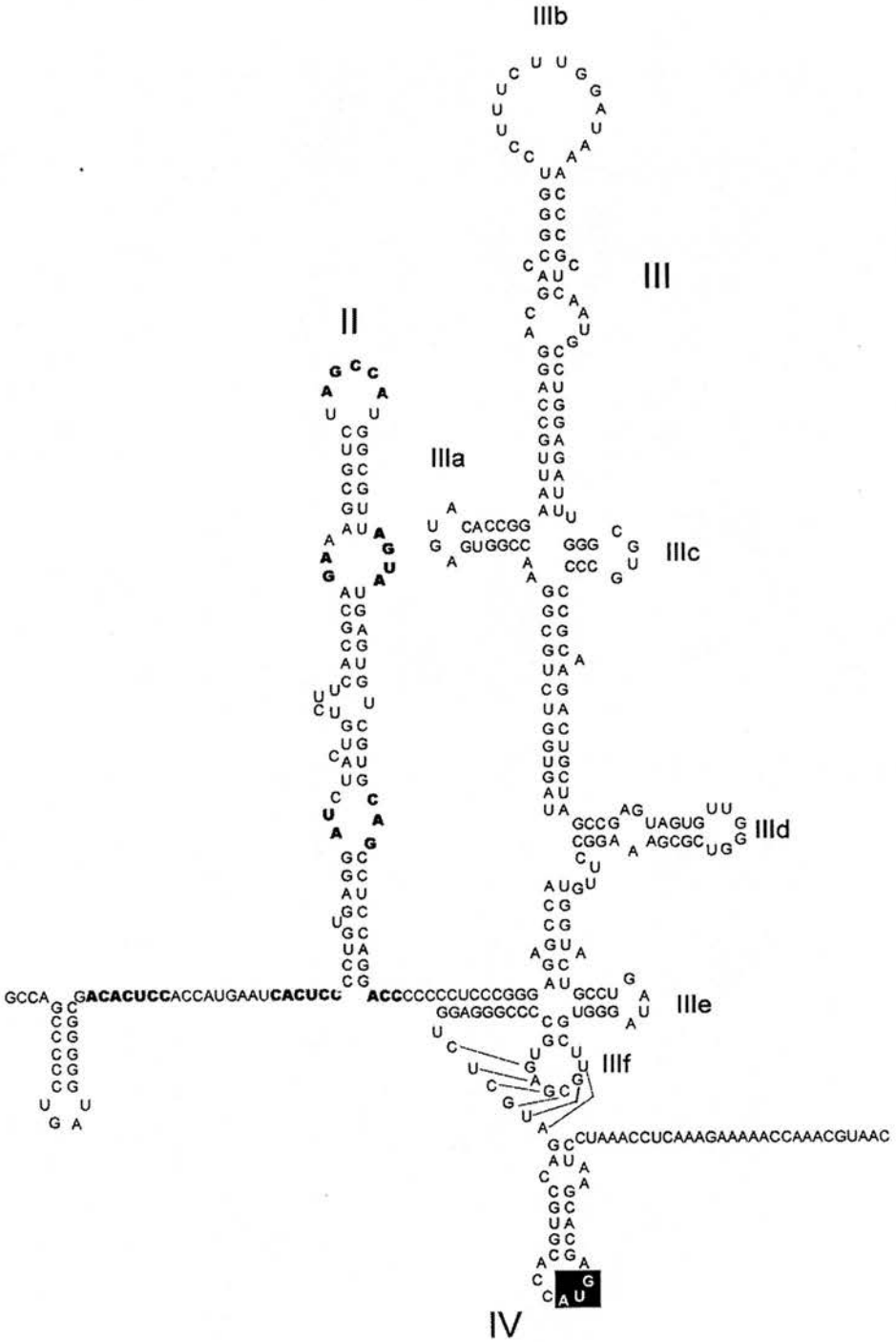


1.5.1 5'UTR

Following characterisation of the 5'UTR region it was shown that polyprotein synthesis is initiated from an AUG start codon at nucleotide 342 (Han *at al.* 1991). The 5'UTR, is predicted to form a stable RNA structures in a similar way to the previously characterised IRESs of picornaviruses (Tsukiyama Kohara *at al.* 1992; Brown *at al.* 1992). The 5'UTR sequence was observed to contain a further four or five (depending on the genotype) potential alternative AUG start codons upstream of position 342 which are believed to be non-functional. The nucleotide sequence of the 5'UTR has been shown to be highly conserved between divergent genotypes (Bukh *at al.* 1992), to such an extent that the primers for the region have been used as the basis of qualitative and quantitative RT-PCR screening between divergent genotypes (1-6) (Smith *at al.* 1995).

The RNA secondary structure of the HCV IRES was modelled by thermodynamic and comparative sequence analysis (Brown *at al.* 1992; Smith *at al.* 1995), and subsequently supported by mutational analysis and ribonuclease mapping (Fig. 1.3) (Honda *at al.* 1996a; Honda *at al.* 1999). The HCV 5'UTR IRES is proposed to be composed of 4 main structural domains (I-IV). The AUG start codon appears to be located in a single stranded loop of the most down stream structure (domain IV), which includes approximately nine nucleotides of the core gene. Mutations which enhance the stability of this structure have been shown to inhibit internally initiated translation (Honda *at al.* 1996a). Some controversy exists as to the exact 3' boundary of the IRES. For example, using a chimeric poliovirus construct that contained the HCV IRES, it was shown that the first 21 nucleotides of the core gene

Figure 1.3. Proposed secondary structure of the HCV IRES. AUG start codon is highlighted in black. Domain names labelled in bold type and start codon highlighted by shaded box. Regions of tertiary interactions highlighted in bold type and by connecting lines. (Adapted from Honda et al 1999).

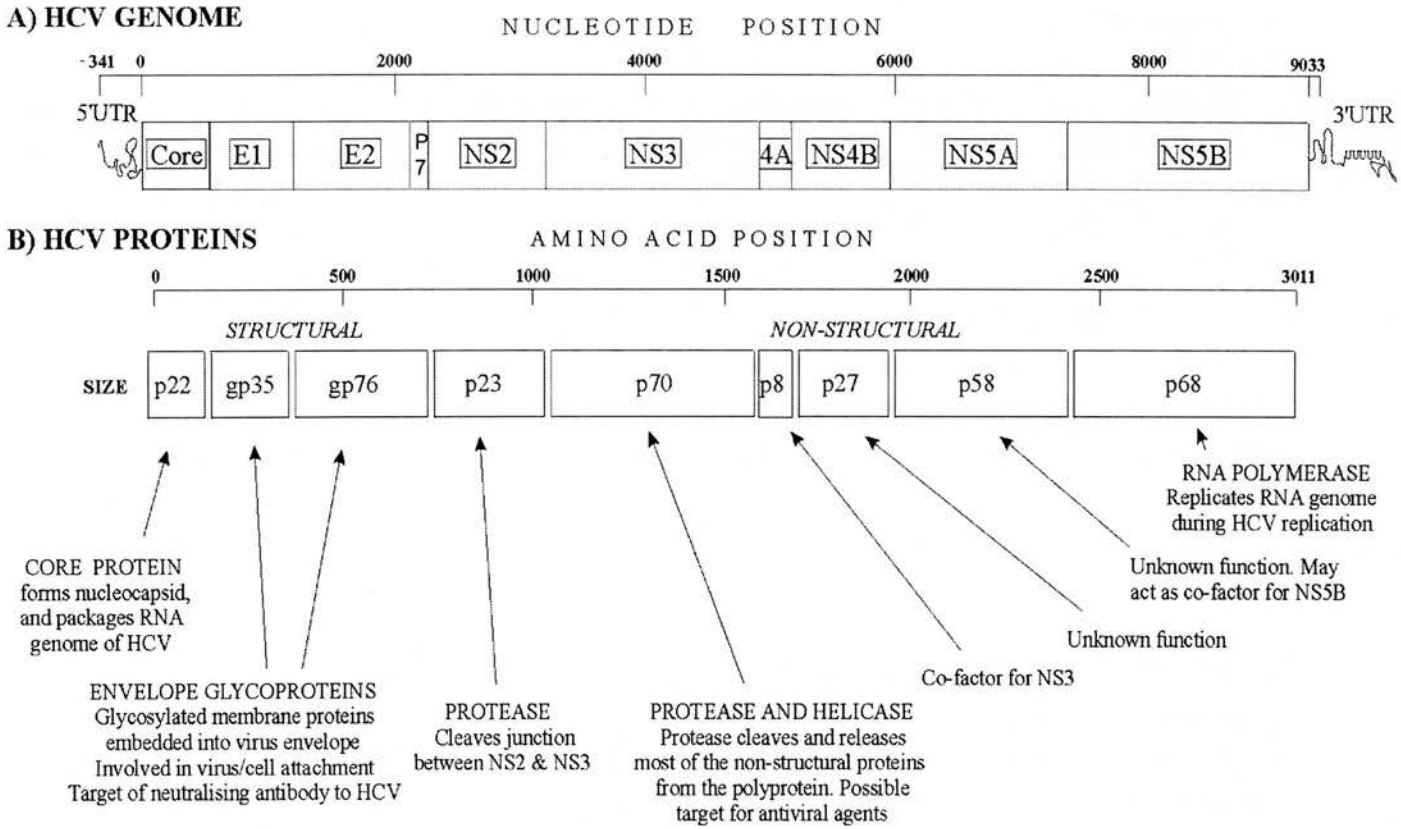


were required for efficient viral replication (Reynolds *et al.* 1995; Lu and Wimmer 1996). However, efficient HCV IRES function has been observed in dicistronic reporter constructs containing only the first AUG start codon of the core gene (Rijnbrand *at al.* 1995; Rijnbrand *at al.* 1996). It has also been proposed that a long range interaction between a region of the core gene sequence and nucleotides 23-38 of the 3'UTR may down-regulate IRES translational efficiency (Kim *at al* 2003). The larger upstream structures (domains II and III) have been shown to be absolutely required for translational activity (Kew 1998; Honda *at al.* 1996b; Rijnbrand *at al.* 1995). The most 5' of the stem loops (domain I) has been observed to have an inhibitory affect on IRES activity, as its deletion appears to up-regulate translation both *in vivo* and *in vitro* (Honda *at al.* 1996b; Reynolds *et al.* 1995; Rijnbrand *at al.* 1995).

1.5.2 CODING REGION

Like other members of the *Flaviviridae* family the nascent viral polyprotein is processed both co- and post-translationally from the N-terminal region by host cell signal peptidases and two viral proteinases. The structural proteins are located in the N-terminal third and non-structural proteins within the C-terminal two thirds of the polyprotein (Fig. 1.4). The structural domain is cleaved by host cell peptidases at the Core/E1/E2/p7/NS2 junction (Grakoui *at al.* 1993b; Mizushima *at al.* 1994; Lin *at al.* 1994; Hijikata *at al.* 1991b; Ralston *at al.* 1993). However, the production of processing intermediates, such as E2-p7-NS2 indicates that not all cleavage of

Figure 1.4. Genomic organisation and gene products of HCV and HGV/GBV-C (Adapted from Simmonds and Smith 1999).



structural proteins is co-translational. Further processing intermediates and truncated protein products have been observed for the structural proteins. For example, the core protein is cleaved from the polyprotein at amino acid 191 (Hijikata *at al.* 1991b; Ralston *at al.* 1993). However, at least 3 truncated core proteins have been detected including a product generated by a secondary cleavage at amino acid 173 which has been isolated from the sera of infected patients, suggesting that the truncated core protein is a component of native viral particles (Yasui *at al.* 1998).

Immunoelectron microscopy and RNA binding assays have been used to confirm the nucleocapsid nature of the core protein. It was shown to form a close association with detergent stripped virion particles recovered from infected hosts and bind RNA encoding the virus structural proteins through a cluster of four basic amino acids at the N-terminus (Takahashi *at al.* 1992) (Santolini *at al.* 1994). A specific interaction between the core protein and the 5'UTR with a resulting suppression of IRES activity has also been reported (Shimoike *at al.* 1999). However, this has been disputed by further research in which no specific interaction was observed (Wang *at al.* 2000). Further putative core protein functions include regulation of cellular transcription, the modulation of cellular transcription, cellular transformation and immunosuppression (Shimoike *at al.* 1999; Wang *at al.* 2000).

The E1 and E2 structural proteins are type 1 membrane proteins which have been shown to contain five/six and eleven N-linked glycosylation sites respectively, in the same relative position as the envelope proteins of flaviviruses and pestiviruses (Choo *at al.* 1991) (Grakoui *at al.* 1993b). The E2 protein has been observed in expression systems extended at its C-terminus to include the small p7 protein (Mizushima *at al.* 1994). Cleavage of E2/p7 and p7/NS2 appears to be a post translational process; the

p7 protein has been implicated in ion channel function (Clarke 1997). The E1 and E2 proteins localise in the endoplasmic reticulum (ER) where they undergo glycosylation and form both non-covalent and di-sulphide bound E1-E2 heterodimers (Flint and McKeating 1999); it is believed that the non-covalently bound dimers represent the mature pre-budding complex (Deleersnyder *at al.* 1997). The cellular attachment and entry mechanism of HCV is still poorly understood. However, E2 is believed to be the main virus receptor mediating virion attachment, as pre-incubation with antisera raised against the E2 protein has been shown to block the attachment process (Flint *at al.* 1999). CD81 was recently identified as a putative cell surface receptor for the E2 protein (Pileri *at al.* 1998), although it has also been speculated that virion entry is via low-density lipoprotein (LDL) receptors. The role of E1 is less clear, although similarities to the envelope fusion proteins of flaviviruses and paramyxoviruses suggest it may play a similar role in HCV membrane fusion.

The E1 and E2 genes are the most genetically divergent regions of the HCV genome. Within E2 there exist hypervariable domains (HVR), including HVR-1, representing the amino-terminal 34 amino acids of the E2 protein (nucleotides 383-414), which is the most variable region of the HCV genome (Kato *at al.* 1992; Weiner *at al.* 1991; Hijikata *at al.* 1991a). The extreme variability of the HVR-1 is believed to play a role in HCV persistence through the development of immunological escape mutations (section 1.8.1) Antibodies raised against this region have been shown to block infection in tissue culture and chimpanzee challenge experiments (Shimizu *at al.* 1996; Farci *at al.* 1996), whilst variability was not observed in a patient with agammaglobulinaemia over a 2.5 year period;

suggesting that the diversity observed in the HVR domains is a result of evolutionary pressure exerted by the host immune system (Kumar *at al.* 1993).

The putative non structural proteins are processed by two virally encoded proteases and as a whole are not expected to be constituents of the mature virion particle but are required for replication of the viral RNA. Processing between the NS2/NS3 boundary is a rapid autolytic event dependant on a NS2-NS3 protease (Silini *at al.* 1993). The proteolytic domain for this activity has been mapped to the C-terminus of NS2 and N-terminus of NS3 (Hijikata *at al.* 1993). It has been proposed that the NS2-NS3 proteolytic domain is a zinc dependant metalloprotease, based upon inhibition with chelating agents such as EDTA (Hijikata *at al.* 1993). NS3 is also involved in the proteolytic processing of all the downstream proteins NS3/4A, NS4A/NS4B, NS4B/NS5A and NS5A/NS5B from a separate serine protease mapped within the C-terminal third of the protein (Grakoui *at al.* 1993a). Analysis of the NS3 protease activity through an *in vitro* transcription-translation system identified a catalytic triad of His (residue 1083), Asp (residue 1107) and Ser (residue 1165) (Bartenschlager *at al.* 1993; Grakoui *at al.* 1993a). In recombinant protein assays, such as vaccinia virus expression systems, substitution of the serine residue was observed to block down-stream polyprotein processing but had no affect on cleavage accoss the NS2/NS3 boundary; further indicating the separate nature of the proteolytic activity at the N-terminal of NS3 (Bartenschlager *at al.* 1993; Grakoui *at al.* 1993a).

NS3 on its own is able to catalyse the proteolytic cleavage of all downstream proteins. However, it was observed that its action is greatly increased in transient and vaccinia virus recombinant expression systems by co-expression of NS4A,

which forms a complex with the N-terminus domain of NS3 and acts as a cofactor; stabilising and localising NS3 to the ER membrane where down-stream proteolytic cleavage is believed to take place (Tanji *at al.* 1995; Failla *at al.* 1995; Wolk *at al.* 2000). The complex crystal formation of the NS3 protease domain and synthetic NS4A cofactor has since been further elucidated by X-ray crystallography (Kim *at al.* 1996).

NS3 is believed to be a multifunctional protein in which amino acid motifs resembling both NTPase and helicase domains have been recognised within the C-terminus of the protein. The NTPase and helicase activity of the NS3 C-terminal domain has been confirmed in recombinant expression systems (Jin,Peterson 1995), and the unwinding properties of the helicase domain have been shown to be optimal in a 3' to 5' orientation. The C-terminus of the protein has also been observed to preferentially bind to both the sense and antisense 3'UTR sequences. Binding to the sense sequence was observed to be dependent on the presence of both the 3' and 5' extremes of the UTR and binding to the negative sense orientation was greatly enhanced by a 3' terminus RNA stem loop structure (Banerjee and Dasgupta 2001).

The NS5B domain of the polyprotein is cleaved into functional domains NS5A and NS5B. The role of NS5A, which is a highly phosphorylated protein, has yet to be fully elucidated although it has been implicated in the modulation of cell cycle genes and resistance of infected cells to the antiviral activity of interferon (IFN). NS5A has been observed in yeast expression systems to bind and inactivate the active site of IFN induced protein kinase (PKR) (Gale *at al.* 1997). The NS5B protein is the most downstream domain of the virus polyprotein and has been shown to possess close sequence identity to the RdRp sequences of many positive stranded RNA viruses,

including other members of the Flaviviridae family (Koonin 1991). In particular the glycine-aspartate-aspartate (GDD) motif, associated with the active site of virus RdRp was observed to be very highly conserved. A number of studies have suggested the HCV RdRp shows little template specificity. However, electrophoretic mobility shift and competition assays have also revealed that recombinant NS5B protein preferentially binds the 3' end of the NS5B protein coding region, between positions 8922 and 9121 (Cheng *at al.* 1999).

1.5.3 3'UTR

The full extent of the HCV 3'UTR has only recently been elucidated. It has a tripartite structure composed of a hetrologous sequence following the ORF stop codon, a poly(U) tract (of variable length) and a highly conserved downstream sequence of approximately 98 nucleotides, known as the 3X' tail (Kolykhalov *at al.* 1996; Tanaka *at al.* 1995; Tanaka *at al.* 1996; Ferri *at al.* 1996). The 3'X tail has since been shown to possess highly conserved RNA stem loop structures which form a terminal "clover-leaf" structure and is essential for virus replication in *vivo* (Kolykhalov *at al.* 1996; Blight and Rice 1997; Smith *at al.* 2002; Yanagi *at al.* 1999; Tanaka *at al.* 1996). It has been shown to bind a number of both viral and cellular proteins, such as NS3 and polypyrimidine tract-binding protein (PTB), which are associated with the viral replication complex (Ito and Lai 1997; Wood *at al.* 2001). It has also been shown, in a recombinant virus expression system, that

inclusion of the 3'X tail of the 3'UTR can result in a three to five fold increase in translation of viral proteins (Ito *at al.* 1998).

1.6 HGV /GBV-C GENOME ORGANISATION

The genome of HGV/GBV-C is a single stranded positive sense RNA molecule approximately 9400 nucleotides in length. The genome and polyprotein organisation of HGV/GBV-C is very similar to that of HCV and other members of the *Flaviviridae* family; with a single ORF, flanked by both 5' and 3'UTRs, which is translated as a polyprotein in a cap independent manner, via an IRES structure situated in the 5'UTR (Simons *at al.* 1996). Although, the genomes of both HCV and HGV/GBV-C are very similar there are several major differences between the two.

1.6.1 5'UTR

The 5'UTR of HGV/GBV-C is longer then that of HCV (524 nt versus 342 nt) and shows limited sequence homology to GBV-A but not HCV or GBV-B (Simons *at al.* 1996). Site specific mutagenesis and protein product sequencing of the translation

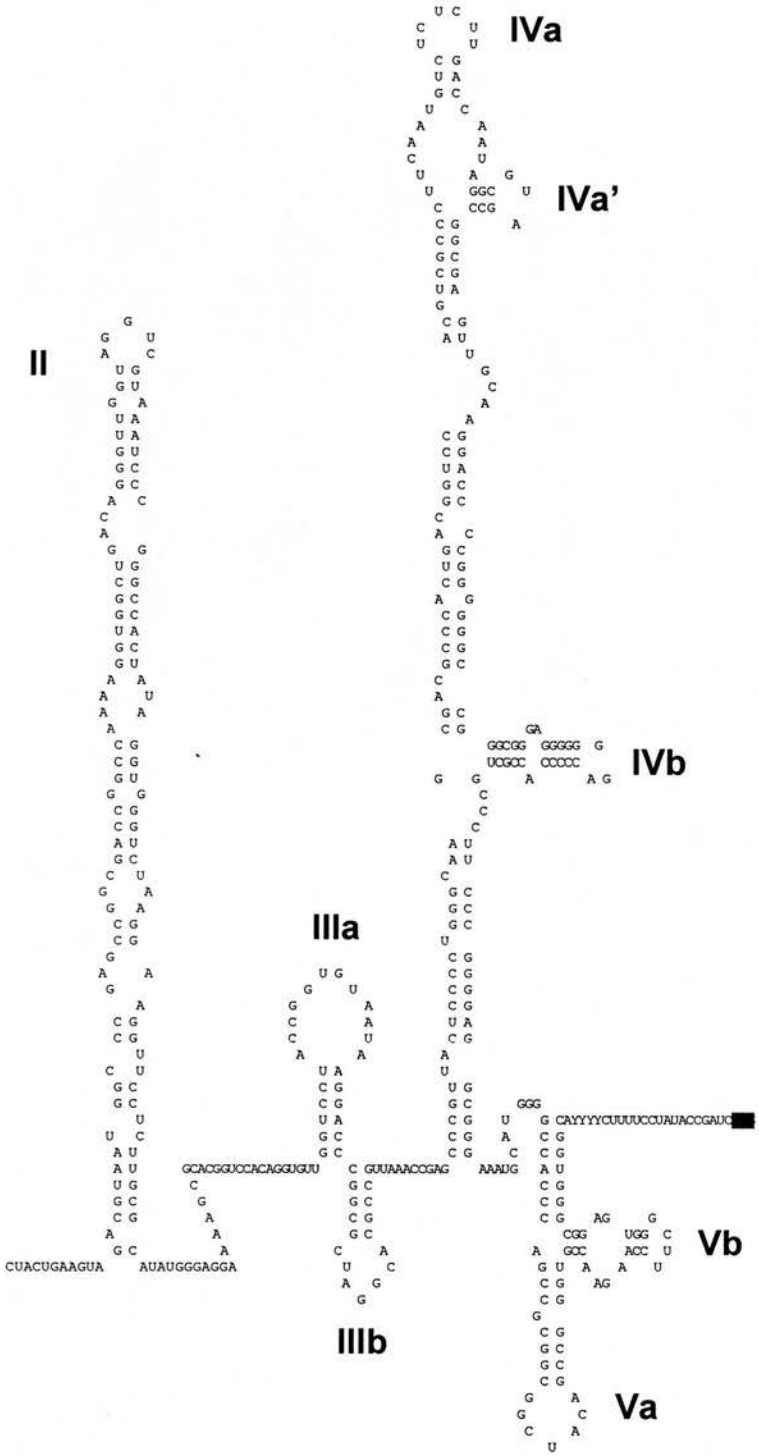
products from a mono-cistronic construct, composed of the 5' region of HGV/GBV-C linked in frame to chloramphenicol acetyltranferase (CAT), indicated that the authentic HGV/GBV-C AUG start codon is located at nucleotide 524; directly upstream of the polyprotein E1 gene (Simons *at al.* 1996). Although a number of other potential AUG start codons exist upstream of this site they have not been observed to be transcriptionally active (Simons *at al.* 1996). Translation of a bi-cistronic construct, containing the HGV/GBV-C 5'UTR, was indicative of the presence of an IRES (Simons *at al.* 1996). However, the activity of the HGV/GBV-C IRES was observed to be much lower (2% to 5%) than that of HCV IRES driven translation.

Based on thermodynamic and phylogenetic analysis the RNA structure of the HGV/GBV-C IRES is predicted to be quite different from that previously modelled for HCV but similar to that of GBV-A (Fig. 1.5) (Brown *at al.* 1992; Smith *at al.* 1995; Simons *at al.* 1996). The IRES is composed of 5 major domains (as opposed to four in HCV) and unlike HCV the AUG start codon is not internal to a stem loop structure but is down stream of a highly conserved oligopyrimidine tract, similar to that critical for IRES driven translation in poliovirus (Simons *at al.* 1996).

1.6.2 CODING REGION

As with HCV, HGV/GBV-C encodes two glyoproteins which are predicted to compose the virus envelope proteins E1 and E2 (Fig. 1.4). Based on putative signal domains it is believed that the E1 and E2 proteins are cleaved from the polyprotein in

Figure 1.5. Proposed secondary structure of the HGV/GBV-C IRES. AUG start codon is highlighted in red. Domain names labelled in bold type (Adapted from Simons et al 1996)



a co-translational manner by host cell peptidases, in a similar manner to those of HCV (Linnen *at al.* 1996; Leary *at al.* 1996). However, several major differences have been recognised between the structure and organisation of the HCV and HGV/GBV-C genomes. Only one potential N-linked glycosylation site has been noted in the E1 protein of HGV/GBV-C and three in the E2 region, which compares to five/six and eleven respectively in HCV (Leary *at al.* 1996; Okamoto *at al.* 1997). A further difference between the viruses are the hypervariable domains such as HVR-1 which are observed in the E2 protein of HCV and are believed to play a role in viral persistence (Kato *at al.* 1992; Weiner *at al.* 1991; Hijikata *at al.* 1991a). Analysis of HGV/GBV-C has shown that, although the E2 encoding region is the most variable across the ORF, the rate of nucleotide and amino acid substitution is much lower than that noted for HCV and no hypervariable regions are present (Katayama *at al.* 1998; Nakao *at al.* 1997). For example, in a study of a patient infected with HGV/GBV-C through a contaminated blood transfusion, approximately the same rate of nucleotide change (0.35%) and only a single amino acid conversion was noted in both the E1 and E2 genes as was observed across the rest of the polyprotein coding region (0.33%) (Nakao *at al.* 1997). In a further study of geographically variant isolates 80% sequence identity was observed across the E1-E2-NS2 region as compared to 50% for HCV (Katayama *at al.* 1998).

A further difference between the genome organisation of HCV and HGV/GBV-C is that HGV/GBV-C does not appear to code for a core or nucleocapsid protein (Linnen *at al.* 1996b; Leary *at al.* 1996) . A number of alternative AUG start codons exist upstream of the recognised start codon at nucleotide 524. However, all have been shown to encode truncated products of variable composition and length or to be

completely absent in many isolates (Okamoto *at al.* 1997). Further, *in vitro* translation experiments using 5'UTR sequences fused in frame with CAT, showed that enzyme activity was abolished when the AUG start codon at position 524 was rendered inactive (Simons *at al.* 1996). In contrast biophysical and electron microscopic characterisations have indicated that the virus may possess a nucleocapsid type structure (Xiang *at al.* 1998). These observations have led to the suggestion that HGV/GBV-C may assemble in a nucleocapsid free virion structure or sequester or utilise host cellular proteins.

The non-structural domains of HGV/GBV-V polyprotein show greater sequence identity to those of HCV. Based on conserved His and Cys residues a zinc dependant metalloprotease has been predicted between the adjacent C-terminal of NS2 and N-terminal of NS3, which is predicted to be responsible for the autolytic cleavage of the NS2/NS3 boundary in a similar mechanism to that of HCV (Linnen *at al.* 1996; Muerhoff *at al.* 1995; Leary *at al.* 1996). Conservation of the amino acid triad His, Asp and Ser, also suggests the presence of a serine protease within the N-terminus of NS3 (Linnen *at al.* 1996; Leary *at al.* 1996). The proteolytic domains and active sites of both enzymes have now been confirmed through site-directed mutagenesis and deletion mapping of the putative motifs, followed by analysis in a baculovirus expression system (Belyaev *at al.* 1998; Khudyakov *at al.* 1997). In this analysis it was shown that the NS3 serine protease is responsible for all downstream cleavages between NS3/NS4A, NS4A/NS4B, NS4B/NS5B and NS5B/NS4B. It was also shown that NS4A was absolutely required as a cofactor for cleavage of NS4B/NS5A and increased the efficiency of the other cleavage events, in a similar way to that observed in HCV (Belyaev *at al.* 1998).

The similarity between the two viruses extends downstream throughout the NS5A and NS5B polyprotein domains, where a number of amino acid sequence motifs have been identified which show close homology to those observed in other viruses of the *Flaviviridae* family. In particular the amino acid motif Gly-Asp-Asp has been identified within the NS5B encoding region, which is indicative of the active site of RdRp. Consequently is it believed that NS5B encodes the HGV/GBV-C polymerase (Linnen *at al.* 1996b; Leary *at al.* 1996).

1.6.3 3'UTR

The sequence of 3'UTR of HGV/GBV-C is highly conserved (Okamoto *at al.* 1997; Katayama *at al.* 1998), with mean pairwise distances of 3.8% to 6.6% between genotypes which is in contrast to 12.8% to 13.4% across the entire genome (Cuceanu *at al.* 2001). Unlike HCV the 3'UTR of HGV/GBV-C is not composed of a tripartite structure and no poly(U) tract has been identified (Okamoto *at al.* 1997; Katayama *at al.* 1998; Cuceanu *at al.* 2001; Xiang *at al.* 2000). Conserved RNA stem loop structures have been observed at the end the 3'UTR of HGV/GBV-C which resemble those predicted in HCV and GBV-B (Okamoto *at al.* 1997; Xiang *at al.* 2000). However, there is no evidence of a conserved third stem loop which makes up the terminal "clover-leaf" structure observed in the other two viruses (Cuceanu *at al.* 2001).

1.7 EPIDEMIOLOGY AND GEOGRAPHICAL DISTRIBUTION

1.7.1 HCV EPIDEMIOLOGY

HCV has been detected at varying levels in populations throughout the world. The main route of transmission for HCV is parenteral exposure to blood. Rates of active infection in healthy blood donors have been estimated to be 0.25% in Australia (Crofts *at al.* 1997); 1.8% in the USA (Alter *at al.* 1999; Alter 1997); 0.69% in France (Aymard *at al.* 1993); 0.66% in Saudi Arabia (Bernvil *at al.* 1994) and 2.8% in the rural Bantu population of the Central African Republic (Fretz *at al.* 1995).

Prior to the development of serological screening assays the main route of transmission was through transfusion of blood and non-inactivated blood products such as factor VIII and IX concentrates. In a study of haemophiliacs who had received factor VIII, prepared from commercial and volunteer plasma, 96.4% (54/56) were later shown to be viraemic for HCV infection by RT-PCR (Jarvis *at al.* 1996). In a further group who, received factor IX, viraemia was observed in 88% (15/17) of individuals (Jarvis *at al.* 1996). All the haemophiliacs who tested negative by RT-

PCR for viraemia were shown to be antibody positive for HCV, indicating past resolved infection.

The risk of exposure through blood transfusion or blood products has been dramatically reduced since the advent of serological screening (Schreiber *at al.* 1996), and the main risk of infection in industrialised countries is now associated with intravenous drug users (IVDUs). Current seroprevalence rates within this risk group are extremely high, ranging from 30% to 90% (Hope *at al.* 2001; Macdonald *at al.* 2000; Taylor *at al.* 2000; Trepo and Pradat 1999). In non-industrialised countries transmission routes are less clear, although the use of non-sterilised medical equipment appears to be a major risk factor (Sanchez *at al.* 2000).

Unusually high seroprevalence has been observed in Egypt of between 20% to 30% in the general population and as high as 50% around the Nile Delta, of which approximately 70% correspond to subtype 4a (Saeed *at al.* 1991; Ray *at al.* 2000; McOmish *at al.* 1994; Simmonds *at al.* 1993). Based on the high predominance of a single subtype, in contrast to the wide diversity of subtypes found in low endemic populations, it is suggested that these high rates of infection are due to a recent spread within Egypt; possibly during a mass campaign, from the 1960s to 70s, to treat schistosomiasis with parenteral treatment (Frank *at al.* 2000).

The frequency of vertical transmission of HCV is comparatively low. For example, in a study of 170 anti-HCV positive women a vertical transmission rate of 2.7% was observed, although a higher rate of 5.4% (2/37) was observed for women co-infected with HIV than in HIV-negative women (3/151) (Ferrero *at al.* 2003). However, all infected children were born to mothers with high levels of viraemia, which is suggestive that many cases may result from unrecognised parenteral

exposure (Ferrero *at al.* 2003). There is evidence that HCV can be sexually transmitted (Kao *at al.* 2000). For example, in a study of 38 spouses of HCV infected patients 7 were found to be viraemic for HCV (Benezra 1993), although the efficiency of sexual transmission is believed to be poor, which was shown in a report that none of 50 partners of HCV viraemic patients had detectable levels of anti-HCV antibody (Bresters *at al.* 1993).

1.7.2 HCV GEOGRAPHICAL DISTRIBUTION

HCV has been shown to group in six major genotypes which exhibit approximately 30% divergence between each other (Simmonds *at al.* 1994a). In Japan and other countries of the Far East, infection is mainly restricted to genotypes 1b, 2a and 2b (Hara *at al.* 1996), whilst in Western Europe and the USA genotypes 1a, 1b, 2b, 2c and 3a predominate although relative frequencies of infection vary. In North and Western Europe the main genotypes are 1a and 3 (Harris *at al.* 1999; Westin *at al.* 1999), in contrast to Eastern Europe where 1b predominates (Naoumov 1999; Stamenkovic *at al.* 2000), and southern Europe where 1b and 2c are most common (Nousbaum *at al.* 1995; Tisminetzky *at al.* 1994; Maggi *at al.* 1999).

In Africa, South East Asia and India the distribution and diversity of genotypes is markedly different from that observed in other regions. Genotypes 3 and 6 predominate in India and South East Asia where numerous divergent subtypes are observed (Simmonds *at al.* 1996). In Western Africa infection is predominantly confined to highly divergent subtypes of genotypes 1 and 2 (Ruggieri *at al.* 1996). In

Central Africa the majority of infection has been shown to be due to highly divergent subtypes of genotype 4 (Fretz *at al.* 1995; Xu *at al.* 1994).

It is believed that areas of high genetic diversity such as West and Central Africa indicate long term evolution of the virus in the host population. The widespread dispersal of a limited number of subtypes amongst specific risk groups in industrialised nations suggest relatively recent introduction of the virus, from areas of endemic infection, into new and efficient transmission networks such as blood transfusion and IVDUs (Nousbaum *at al.* 1995; Westin *at al.* 1999; Yu *at al.* 2001; Tisminetzky *at al.* 1994).

1.7.3 HGV/GBV-C EPIDEMIOLOGY

HGV/GBV-C has been shown to be more prevalent worldwide in healthy population groups than HCV. Levels of active viraemia measured by RT-PCR have been measured in Scotland at 3.2% (Jarvis *at al.* 1996); France 3.4% (Mercier *at al.* 1999); Japan 1% to 2.9% (Zhang *at al.* 1998) and 5% in Thailand (Poovorawan *at al.* 1998). However, these figures underestimate rates of past exposure to HGV/GBV-C as measured by the presence of anti-HGV/GBV-C antibody, which is rarely observed in viraemic patients. Based on combined serological and RT-PCR analysis past and present exposure has been shown to be remarkably high in healthy blood donor populations with rates of 12.9% in France (Mercier *at al.* 1999); 16% in Spain (Tacke *at al.* 1997); 9% in Sao Paulo Brazil (Bassit *at al.* 1998), 13.3% in Australia and 10.4% in Germany (Nubling *at al.* 1997).

As with HCV, parenteral exposure from contaminated blood or blood products and intravenous drug use, is a main risk factor associated with transmission of HGV/GBV-C. For example, present or past exposure in IVDUs in Germany has been shown to be as high as 74.4% (Nubling *at al.* 1997); 88.6% in Australia (Hyland *at al.* 1998) and 52% to 73% in Spain (Tacke *at al.* 1997). Viraemia in multiply transfused patents has been observed at increased frequencies of 30% (Chen *at al.* 1997) and 32.6% (Poovorawan *at al.* 1998). High rates of viraemia have also been observed in patients who have received non-virus inactivated blood products. In a study of 95 Scottish haemophiliacs, 14% where shown to be positive for HGV/GBV-C viraeimia (Jarvis *at al.* 1996). However, this compares with 83% who where shown to be positive for HCV replication, despite similarly high levels of contamination of blood products, suggesting that infection which HGV is less likely to establish a chronic infection than HCV (Jarvis *at al.* 1996).

Unlike HCV, high numbers of sexual partners and vertical transmission from mother to baby have been shown to be associated with a high risk of HGV/GBV-C transmission. A study of homosexuals in Germany showed a frequency of 30.2% past exposure to the virus (Nubling *at al.* 1997) and past and present exposure has been measured at approximately 45% in homosexual men and prostitutes (Scallan *at al.* 1998). A number of studies have presented results consistent with vertical transmission which has been shown to be more frequent for HGV/GBV-C than HCV and HIV (Feucht *at al.* 1996; Zanetti *at al.* 1998).

1.7.4 HGV/GBV-C GEOGRAPHICAL DISTRIBUTION

HGV/GBV-C has been divided into four main genotypes based on comparison of complete genome sequence (Smith *at al.* 2000). The maximum divergence between genotypes is approximately 13% at the nucleotide levels and 4% of amino acid residues, compared to over 30% nucleotide divergence across the length of the HCV genome. The four genotypes show a distinct geographical distribution (Smith *at al.* 2000; Kondo *at al.* 1997); genotype 2 is predominant in Western Europe, America, India and Northern Africa, genotype 4 in South East Asia and genotype 1 in Sub-Saharan Africa which also exhibits the greatest observed diversity, particularly in the 5'UTR (Tucker *at al.* 1999; Sathar *at al.* 1999). Genotype 3 predominates in Northern Asia and the native inhabitants of North and South America (Konomi *at al.* 1999; GonzalezPerez *at al.* 1997; Tanaka *at al.* 1998b; Konomi *at al.* 1999). It has been proposed that the major observed geographical distribution of HGV/GBV-C is a consequence of early *Homo sapien sapiens* migration (Smith *at al.* 2000).

1.8 DISEASE ASSOCIATIONS

1.8.1 HCV

Acute or primary HCV infection is generally asymptomatic, or associated with a mild illness. However, in at least 50% of infected individuals, HCV establishes a persistent infection of the hepatocytes which is associated with chronic inflammation

of the liver (Alter *at al.* 1992). The average period of incubation from exposure to the onset of acute symptoms is approximately 6 to 8 weeks, with the appearance of anti-HCV antibodies, detectible by ELISA, between weeks 8 and 9. The time between infection and detectible levels of anti-HCV antibodies has been termed the "window period" during which serological screening of blood donors is ineffective. However, viraemia has been detected within 1 week of experimental infection or transfusion (Shimizu *at al.* 1990; Farci *at al.* 1991). Consequently RT-PCR may be used to detect HCV RNA in blood donors, reducing the risk from window period blood donations. Acute HCV infection is generally subclinical. When symptoms are observed they are commonly indistinguishable from those due to other forms of acute viral hepatitis and include jaundice, fatigue, anorexia and abdominal discomfort. Unique to HCV infection is the extreme likelihood of the virus causing a chronic infection and the rarity of progression to fulminant liver failure (Alter *at al.* 1992), although such progression has been documented in a few cases (Yoshiba *at al.* 1994).

Most of the serious chronic liver diseases associated with HCV infection are due to cytopathic infection of hepatocytes and the associated inflammatory response. However, there is also evidence that many of the pathological aspects of HCV infection are due to a host autoimmune response, which is also believed to be responsible for a number of extrahepatic symptoms including cryoglobulinaemia (Ballardini *at al.* 1995; Johnson *at al.* 1993). Progression to clinically significant liver disease is generally slow and insidious, with symptoms commonly absent or non-specific over long periods of time. Intermittent elevated ALT levels are observed in approximately 70% of patients reflecting the damage suffered by

hepatocytes (Di Bisceglie 1998). Histological characteristics of chronic infection include the appearance of aggregates in the portal tract, inflammatory infiltrates in the parenchyma, fibrosis, necrosis in periportal regions and the accumulation of lipid vacuoles in the cytoplasm of infected hepatocytes Over a period of years to decades the inflammatory process may progress to cirrhosis followed by hepatocellular carcinoma (HCC) in a further 20% to 25% of patients (Benvegnu *at al.* 1997; Tsukuma *at al.* 1993). In order to make an assessment of a chronically infected patient's likely disease progression, hepatitis activity is graded over time based on the levels of necroinflammatory activity and fibrosis.

A number of factors have been implicated in the rate and likelihood of chronic HCV infection progressing to cirrhosis and/or HCC. These include a history of alcohol abuse (Ostapowicz *at al.* 1998), infection with other hepatitis viruses such as HBV or HAV (Chiaramonte *at al.* 1999), route of inoculation, host age and infecting virus genotypes. For example, infection with genotype 1b has been documented to result in a more aggressive disease course resulting in higher incidences of cirrhosis and HCC (Seeff 1998), although more recent studies have failed to confirm this.

Infection with HCV elicits both T-cell mediated and humoral immune responses and may be responsible for clearance of the virus in individuals in which virus infection does not progress to chronicity. For example, in two recent studies a strong association between a vigorous and sustained $CD4^+$ T-cell response and a self limiting course of acute infection was observed (Diepolder *at al.* 1995; Missale *at al.* 1996). The observed T-cell responses were directed against a number of viral antigens (Core, E2, NS3, NS4 and NS5), although in the majority of individuals in which viral clearance was observed the response against NS3 was strongest and

detected most consistently (Diepolder *at al.* 1995; Missale *at al.* 1996; Diepolder *at al.* 1997).

Despite both humoral and cellular immune responses, spontaneous clearance of virus during chronic infection is rare and the majority of acutely infected patients progress to chronic infection. It has also been shown that HCV infection does not illicit a protective immune response (Farci *at al.* 1992); highlighted by the fact that following a self limiting infection or during chronic disease re-infection with a homologous or heterologous serotype is possible. HCV is believed to escape immune surveillance and clearance through a number of mechanisms. A major factor is believed to be viral heterogeneity, in which the host immune system exerts evolutionary pressure leading to an increasingly complex population that can elude the immune attack and result in a persistent infection (Chang *at al.* 1997; Wang and Eckels 1999). For example, in a recent study of acutely infected post transfusion patients those which resolved the infection exhibited relatively little diversity in hypervariable region 1 (HVR1), of the envelope protein E2, compared to those with chronic infection in which increasing levels of sequence diversity were observed (Farci and Purcell 2000)

Further mechanisms of immune evasion are believed to include association of virion particles with plasma lipoproteins (Thomssen *at al.* 1993); impairment of dendritic antigen presenting cell function (Bain *at al.* 2001) and a direct interaction between viral proteins and the host immune response. The NS5A viral protein has been shown to directly interact with and inactivate an interferon (IFN) induced protein kinase, PKR, which is a mediator of IFN-induced antiviral resistance (Enomoto *at al.* 1995; Gale *at al.* 1997). Viral amino acid substitutions within NS5A

have also been implicated in a reduced response to treatment of chronic genotype 1b infection with IFN-α, which is the most widely accepted form of therapy for chronic HCV infection (Fukuda *at al.* 1998; Gale *at al.* 1998; Murashima *at al.* 1999).

1.8.2 HGV/GBV-C

The detection of HGV/GBV-C in sera from patients with non A-E hepatitis aetiology initially suggested that HGV/GBV-C might be responsible for some cases of non A-E hepatitis (Linnen *at al.* 1996; Yoshiba *at al.* 1995). However, subsequent investigations of HGV/GBV-C viraemia and acute or chronic hepatitis have failed to provide convincing causal evidence that HGV/GBV-C infection is associated with any pathological process. In the United States HGV/GBV-C RNA was found in 9% of patients with acute non A-E hepatitis, none of which developed chronic hepatitis over a follow up period of nine years, despite persistent viraemia in 75% of patients (Alter *at al.* 1997b). HGV/GBV-C infection has also been implicated in the occurrence of fulminant hepatitis (Yoshiba *at al.* 1995; Heringlake *at al.* 1996). A strain of HGV/GBV-C with a specific sequence motif domain within the NS3 region was isolated from eleven patients with fulminant hepatic failure (Heringlake *at al.* 1996). However, patients with fulminant hepatitis receive multiple transfusions so it was not possible to assess the role of HGV/GBV-C in the disease aetiology of the patients. In a further study fourteen patients undergoing liver

transplant operations, due to unexplained fulminant hepatitis, were tested for HGV/GBV-C RNA (Moaven *at al.* 1997). All fourteen patients tested negative for HGV/GBV-C RNA pre-operation, three patients died waiting for a suitable donor and eight of the eleven surviving patients tested positive for viral RNA post-operation. This was compared to a control group of twenty one post transplant patients, five of which were positive for viral RNA. These results and others suggest that the high rates of HGV/GBV-C infection observed in fulminant hepatitis are not the aetiological agent but are more likely due to multiple transfusions.

A number of studies have reported that the rate of HGV/GBV-C viraemia is independent of patient serum ALT levels and shows no correlation to chronic hepatitis (Linnen *at al.* 1996; Bosmans *at al.* 1997; Sarrazin *at al.* 1997). Further, no significant correlation has been observed between HGV/GBV-C infection and histological features of chronic hepatitis, such as fibrosis or inflammation (Haagsma *at al.* 1997; Sarrazin *at al.* 1997; Bralet *at al.* 1997). No link has been observed between co-infection of HCV infected patients with HGV/GBV-C and the severity or occurrence of clinical progression (Alter *at al.* 1997b; Alter *at al.* 1997a; Bralet *at al.* 1997; Sarrazin *at al.* 1997). For example in a study of 105 HCV viraemic patients, no statistical difference was observed in the incidence of cirrhosis or level of hepatic fibrosis, between those solely infected with HCV and those co-infected with HGV/GBV-C (Bralet *at al.* 1997).

Based on the presence of antisense RNA replication intermediates HGV/GBV-C replication has been detected in hepatocytes of the liver (Saito *at al.* 1997; Fan *at al.* 1999). However, this has not been widely reproducible and most groups have observed higher levels of replication in plasma, peripheral blood mononuclear cells

(PBMCs), bone marrow and spleen tissue; suggesting that HGV/GBV-C is predominantly a lymphotropic virus (Pessoa *at al.* 1998; Laras *at al.* 1999; Laskus *at al.* 1998; Fan *at al.* 1999; Kao *at al.* 1999; Tucker *at al.* 2000).

As with HCV, HGV/GBV-C is able to setup a long term persistent infection. However, levels of HGV/GBV-C persistence are lower (approximately 20%) than those observed for HCV. This difference may be due to the presence of protective antibody against the E2 envelope protein, which is usually associated with past exposure and clearance of the virus and is rarely observed in viraemic patients (Feucht *at al.* 1997; Thomas *at al.* 1998). This contrast to HCV may be due to higher levels of sequence homogeneity observed in HGV/GBV-C, in particular the envelope proteins of HGV/GBV-C do not contain hypervariable regions (Katayama *at al.* 1998; Nakao *at al.* 1997), which are believed to play a role in the immune escape of HCV. Based on structural homology the NS5A protein of HGV/GBV-C may play an analogous role to that of HCV, in which an interaction with host kinase PKR modulates the host antiviral IFN response.

In summary, no apparent disease burden has been associated with HGV/GBV-C infection, hepatic or non-hepatic and it is currently believed that the main site of virus replication is outside the liver and may be related to PBMCs. Also the mechanism by which HGV/GBV-C persistent infection is maintained is poorly understood and requires further characterisation.

1.9 STRUCTURAL CONSTRAINTS ON HCV AND HGV/GBV-C VIRUS EVOLUTION

1.9.1 HCV

HCV can be divided into at least six major genotypes differing from one another by approximately 30% at the nucleotide level (Simmonds *at al.* 1994a). A number of studies have attempted to reconstruct the origin of this sequence diversity based on longitudinal studies of infection in one or more individuals with known times of infection (Ogata *at al.* 1991; Okamoto *at al.* 1992; Abe *at al.* 1992). Sequence analysis of virus isolated from a patient infected in 1977 was compared with that isolated from the same patient thirteen years later in 1990 (Ogata *at al.* 1991). The two isolates differed at 2.5% of nucleotides across the complete genome resulting in a rate of nucleotide substitution of 1.92×10^{-3} per site per year. In a further study HCV nucleotide variability was assessed between two viruses, isolated 8.2 years apart, from an experimentally infected chimpanzee (Okamoto *at al.* 1992). Across the complete genome a 1.18% difference in nucleotide sequence was observed giving a rate of substitution of 1.44×10^{-3} per site per year. In a further study a slower rate of nucleotide substitution (0.9×10^{-3} per site per year), was observed within the structural region of a virus isolated 9 years post experimental inoculation in a chimpanzee and that of the original virus sequence (Abe *at al.* 1992b). Extrapolation of the average substitution rates from these studies results in an estimated time of

origin for the virus subtypes of 135-286 years ago and genotypes of 230-480 years ago (Simmonds and Smith 1999a).

However, in both the Ogata and Okamoto studies differing degrees of sequence variation were observed between different subgenomic regions (Abe *at al.* 1992; Ogata *at al.* 1991). For example, 0.7% sequence divergence was observed between the sequences of the 5'UTR regions, isolated from the same patient thirteen years apart, compared to a 28.2% nucleotide divergence between the envelope genes (Ogata *at al.* 1991). Such inconsistencies in the frequency of nucleotide substitutions across the genome mean that divergence in highly variable regions may become saturated over a relatively short time span, leading to inconsistencies when determining overall rates of substitution.

A further compounding factor, which may result in false estimates of rates and times of sequence divergence, is uncertainty within longitudinal studies over the relationship of compared sequences (Simmonds and Smith 1999a). For example, primary infection may be initiated with many virus particles, thus a divergent virus sequence detected after some time may have been present but undetected in the initial inoculum, leading to an overestimation of the rate of sequence change. In a previously described study (Okamoto *at al.* 1992), a chimpanzee was infected with serum pooled from other chimpanzees originally infected from a chronically infected blood donor nine years previously. On analysis a third as much variation was observed in sequences cloned from the initial patient sera as was observed over time. Further, equal levels of diversity were observed between different clones isolated from the final chimpanzee infection as was noted between the initial patient and chimpanzee sequences, which were isolated several years apart (Okamoto *at al.*

1992). Consequently, some of the variation observed over time, in this and other studies, may represent turnover of sequences present at the initial time of transmission and so actual rates of nucleotide substitution may be lower than previously estimated (Smith *at al.* 1997b).

In order to overcome this apparent sampling bias a further investigation was undertaken in which the rates of substitution in both the E1 and NS5B coding regions were measured in multiple recipients from the same infectious source (Smith *at al.* 1997b). The infectious source had been identified as anti-D immunoglobulin and diversity was measured in 26 individuals, 17 years post infection. After analysis approximately half as much sequence divergence was observed between recipients compared to between recipients and sequence isolated from the infectious serum, for both the E1 and NS5B genes. This is consistent with the shorter period of divergence from the source (17 years) than between recipients (34 years). In order to further account for potential sampling bias, individual virus genomes were sequenced from the initial infectious serum after obtaining PCR product through limiting dilution PCR. The overall rate of substitution between the recipient samples was calculated to be 0.41×10^{-3} per nucleotide per year for NS5B and 0.74×10^{-3} for E1 (Smith *at al.* 1997b). Using these lower rates of substitution times of origin of 260-280 years ago for virus subtypes and 400-570 years ago for virus genotypes has been estimated (Simmonds and Smith 1999a).

Comparison of complete genome sequences for different virus genotypes has revealed that variability at synonymous sites, which are presumed to be evolutionarily neutral, is not constant across the genome (Ina *at al.* 1994; Smith *at al.* 1997b; Smith and Simmonds 1997). A study of approximately 100

epidemiologically unlinked complete genome sequences (genotypes 1a to 6a), revealed suppression of synonymous variability towards both the 3' and 5' ends of the polyprotein encoding sequence (Tuplin *at al.* 2002). The greatest constraints in synonymous variability were observed across the core gene and NS5B region, in which an approximately three fold reduction in variability compared to the rest of the genome, was observed (Tuplin *at al.* 2002). Several different explanations can be advanced for this bias.

Suppression of synonymous variability could be due to the presence of alternative overlapping reading frames, although this has been shown to be unlikely. For example, an alternative reading frame has been proposed within the core gene from the observation that the +1 reading frame is open between codons 124 to 163 (Ina *at al.* 1994). However, no AUG start codon is present and further analysis has shown that in approximately 10% of sequences this short ORF contains at least one stop codon (Bukh *at al.* 1994). In a recent study expression of the +1 ORF from a vaccinia virus construct was shown to require mutations within the core gene to allow ribosomal frame shifting and the recombinant protein was observed to be extremely unstable (Roussel *at al.* 2003). Further, in a set of 52 unlinked sequences only 23% of amino acid positions are conserved between positions 2 to 160 in the +1 reading frame with a similar proportion in the +2 reading frame (Smith and Simmonds 1997). The +2 reading frame also contains numerous stop codons and the longest ORF is only 33 codons in length, between codons 121 to 154. All three antisense reading frames also encode numerous stop codons and few AUG start codons. Similar results have been observed for the NS5B encoding region (Smith and Simmonds 1997).

A further possibility is that suppression of synonymous variability results from biased codon usage within these regions. However, base composition in the first, second and third codon positions has been shown to be consistent across the genome of HCV, with no evidence of differing codon usage within regions in which synonymous variability is constrained (Tuplin *at al.* 2002). Suppression on synonymous variability may also result from biased dinucleotide frequencies which limit codon choice. However, this has been discounted as for 14 of the possible 16 dinucleotides little or no difference from expected frequencies has been observed. The dinucleotide CG is slightly under represented and UG overrepresented, although no correlation has been observed between frequencies of either dinucleotide and regions in which synonymous variability is constrained (Tuplin *at al.* 2002).

Reduction in sequence diversity at synonymous sites may result from constraints on sequence change that arise from the formation of RNA stem loop structures that influence virus phenotype. It has been shown that the first eight nucleotides of the core gene take part in base pairing with upstream nucleotides at the 3' end of 5'UTR, forming the terminal domain of the HCV IRES (Brown *at al.* 1992; Smith *at al.* 1995). At least two further stem loop structures have been predicted downstream of this structure by thermodynamic minimisation between nucleotides 47-84 and 87-167. Two RNA structures have also been predicted towards the 3' extreme of the NS5B coding regions between nucleotide positions 8979-9011 and 8917-8976 (Han and Houghton 1992; Smith and Simmonds 1997; Kolykhalov *at al.* 1996; Tanaka *at al.* 1996).

The presence of functionally important RNA structure would be expected to confer constraints on nucleotide variation at synonymous sites due to the requirement to

maintain base pairing. However, reduction in synonymous variability is observed across the complete length of the core gene and across approximately the final 2000 nucleotides of the polyprotein coding region (Tuplin *at al.* 2002). Although, the RNA structures discussed above may account for a proportion of the observed synonymous bias they are not extensive enough to account for the scale of sequence constraint observed. A further indication that RNA structure may be more extensive, is the observation that covariant substitutions have been shown to cluster in regions of extreme synonymous bias within HCV (Tuplin *at al.* 2002; Smith and Simmonds 1997). Although, a number of these clusters are accounted for by the structures discussed most are not. In particular at least seven clusters of covariant sites have been observed across the NS5B region whilst only two limited structures were predicted in this region. Consequently, phylogenetic evidence is suggestive of extensive evolutionarily conserved RNA structures within both the core gene and NS5B region of HCV.

1.9.2 HGV/GBV-C

Based on analysis of complete genome sequences HGV/GBV-C has been divided into four major genotypes with distinct geographical locations and 11% nucleotide difference between the most divergent genotypes (Smith *at al.* 2000). However, these genotypes are not consistently reproduced by analysis of the 5'UTR (Smith *at al.* 1997a), or subgenomic regions (Khudyakov *at al.* 1997), with the exception of the E2 gene (Smith *at al.* 2000). The inconsistent phylogenetic relationships observed

between HGV/GBV-C subgenomic regions is in sharp contrast to HCV, in which phylogenetic analysis of a number of subgenomic regions has been shown to reconstruct the relationship observed between complete genome sequences (Simmonds *at al.* 1994b).

In order to reconstruct the relationship between genotypes of HGV/GBV-C and estimate times of divergence a number of studies have attempted to define rates of sequence change for the virus. For example, in a study of a patient on maintenance haemodialysis who had been infected with HGV/GBV-C through transfusion the sequence of virus isolated from the patient at seroconversion was compared with that isolated 8.4 years later (Nakao *at al.* 1997). The two isolates differed at 0.33% of nucleotide sites and 0.18% of amino acid residues giving a rate of nucleotide sequence change of 0.4×10^{-3} per site per year. In a further study sequence divergence was measured across nucleotides of the NS4/NS5 region over 2.5 years in an individual with community acquired HGV/GBV-C (Khudyakov *at al.* 1997). A higher rate of substitution over time was noted in this study, with a 1% nucleotide difference between the two isolates giving a rate of sequence change of 1.7×10^{-3} per site per year. These observed rates of evolution are consistent with those observed for HCV during chronic infection ($0.4–1.9\times10^{-3}$ (Ogata *at al.* 1991; Okamoto *at al.* 1992; Abe *at al.* 1992). Based on these rates of annual substitution it has been estimated that the current diversity observed between the most divergent HGV/GBV-C genomes would have required between 120 and 500 years to develop (Simmonds and Smith 1999a).

Rates of short term sequence change equivalent to HCV and a recent divergence of genotypes is inconsistent with several other lines of evidence which suggest that

HGV/GBV-C may have always infected the human population through coevolution with its primate hosts (Simmonds and Smith 1999b). The virus is widely geographically distributed with rates of active or past infection ranging from 5% to 15% and distinct phylogenetic groupings found in Africa, Asia, and North America/Europe (Smith *at al.* 2000; Katayama *at al.* 1998). Viruses within these groupings have been found in isolated populations in Africa, Papua New Guinea and South and Central America as well as highly populated urban areas, which is consistent with long term endemic infection. Although, HGV/GBV-C infection is often chronic and associated with high levels of viraemia no hepatic or non-hepatic disease association has been identified, which is consistent with long term co-evolution with its host. The geographical distribution of HGV/GBV-C genotypes is also consistent with ancient human migrations. For example, genotype 3 has been isolated, almost solely, from populations in the Far East and native inhabitants of North and South America, which is consistent with infection in the human population prior to migration across the Bering land bridge and subsequent colonisation of the Americas, between approximately 15-35 000 years ago. In contrast sequences isolated in Europe, Western India and North Africa are usually genotype 2. Greatest sequence diversity is observed in genotype 1, which is confined to sub-Saharan Africa and exhibits the most extreme variation observed in 15 sequences isolated from Pygmies and Bantu in Central Africa (Tanaka *at al.* 1998a).

Recently a virus closely related to HGV/GBV-C has been isolated from non-captive chimpanzee populations (*Pan troglodytes*, *troglodytes* and *verus* subspecies) from the Central and Eastern African countries of Cameroon and Nigeria. The isolate has been described as HGV/GBV-C_{CPZ} and based on analysis of the 5'UTR,

NS5 and NS3 regions was shown to be less divergent from human sequences (27% nucleotide divergence; 15% amino acid divergence) than those recovered from GBV-A infection in new world primates (Adams *at al.* 1998). HGV/GBV-C_{CPZ} sequences recovered from different subspecies of chimpanzee were shown to be more diverse than variants of HGV/GBV-C isolated from the human population. 25% nucleotide sequence divergence in the 5'UTR and 20% divergence in the NS5 region (9.5% amino acid) between virus isolates recovered from the *troglodytes* and *verus* subspecies compared with 7.4% and 10.4% (1.9% amino acid) divergence between HGV/GBV-C isolated from the human population (Adams *at al.* 1998). The greater divergence observed within the chimpanzee subtype is consistent with the greater nuclear genetic diversity of chimpanzees when compared to humans. This is presumed to be due to the maintenance of chimpanzee population sizes in the past and the relatively recent expansion of limited numbers of modern humans from Africa. Even greater divergence is observed within the HGV/GBV-C homologue (GBV-A) isolated from New World primates, with 42% nucleotide sequence (38% amino acid) divergence within the NS5 and 25% divergence within GBV-A sequences isolated from different New World primates, which again mirrors the host evolutionary relationships divergence times (Simmonds and Smith 1999b; Charrel *at al.* 1999). These observations are consistent with coevolution of the virus with its primate host. A comparison of the phylogenetic relationships observed among HGV/GBV-C and GBV-A isolates and the phylogeny of their primate hosts reveals striking similarities between the branching pattern of the viruses and those of the primates. From this comparative study a rate of sequence change of 3×10^{-7} per site per year has been inferred (Charrel *at al.* 1999). Comparable results were obtained

in a similar study, in which rates of sequence divergence of 5.6×10^{-6} and 3.2×10^{-6} nucleotide and synonymous nucleotide substitutions per site per year respectively, were inferred (Simmonds and Smith 1999b).

In the studies of chronically infected individuals discussed previously, HGV/GBV-C was shown to have a short term rate of sequence divergence, over 8.4 years, comparable to that of other RNA viruses such as HCV (Nakao *at al.* 1997), which is incongruent both with early infection of the human population, before the migration of modern humans out of Africa, and the lack of diversity observed. The short term rate of sequence change is also inconsistent with the relationship between related viruses and cospeciation with primate hosts. If the virus did not coevolve with its host then the geographical spread would have to be due to recent zoonotic events, which is inconsistent with the geographical spread if the virus genotypes and the relationship between related viruses and their hosts. This suggests that constraints on sequence variability exist, which restrict the accumulation on nucleotide substitutions in HGV/GBV-C, limiting diversity over time (Simmonds and Smith 1999b).

Further evidence for this has been highlighted by the analysis of the rate of synonymous substitutions, which are not limited by constraints on the encoded amino acid sequence. In a study of 17 epidemiologically unlinked HGV/GBV-C complete genome sequences (genotypes 1-3) 23% (622) of codons were invariable at their synonymous sites; in a second dataset of 17 genotype 3 sequences the bias was even greater with 30% of codons invariable at their synonymous sites (Simmonds and Smith 1999b). As with HCV the possibility that these rates were due to unequal codon usage or base compositions biases was discounted and there was no evidence of overlapping reading frames in either the sense or antisense orientation (Simmonds

and Smith 1999b). However, the degree of bias in synonymous substitutions was more extreme than that observed for HCV and observed along the length of the HGV/GBV-C genome; not restricted to the 5' and 3' regions as was observed in HCV.

In the same study multiple covariant substitutions were observed along the length of the genome, with greater clustering in regions which were spatially associated with extreme reductions in synonymous variability, such as the NS5, NS3 and E2 protein encoding regions (Simmonds and Smith 1999b) (Cuceanu *at al.* 2001). Consequently it was suggested that the genome of HGV/GBV-C may contain evolutionarily conserved stem loop structures which would limit the rate of synonymous and non-synonymous sequence variability, due to the need to maintain internal base pairing. Based on covariant substitutions a number of RNA structures were predicted along the length of the genome and were again shown to align which regions of synonymous nucleotide bias (Simmonds and Smith 1999b). One of these structures, which was located in the E1 gene, was shown to be thermodynamically stable and was mapped.

1.9.3 COMPARISONS BETWEEN HCV AND HGV/GBV-C

An intriguing finding from these comparisons is the extent to which the characteristics of sequence evolution differ so markedly between HCV and HGV/GBV-C. Although, both sequences show comparable short term rates of sequence change the long term rate of divergence is severely restricted in

HGV/GBV-C. Both viruses exhibit constraints in the rate of substitution at synonymous sites and show geographically distinct genotypes. However, reductions in synonymous variability are more pronounced and divergence between genotypes far more restricted in HGV/GBV-C. These biases and differences may be accounted for by the presence of evolutionarily constrained RNA structure in both viruses which is more extensive in HGV/GBV-C than HCV.

Whilst the reason for this difference, given the close relationship of the two viruses, is not understood it may account for the lack of sequence divergence between genotypes of HGV/GBV-C. These findings may also account for the fact that genotypes of HCV can be resolved by subgenomic fragments as short as 222 nucleotides whilst complete genomes are generally required for HGV/GBV-C.

A further difference between the viruses is the extent of sequence conservation within the 5'UTR regions. The 5'UTR of HCV is the most strongly conserved region across the genome with divergence between genotypes of between 2.5 and 11%. In HGV/GBV-C the 5'UTR exhibits approximately the same level of conservation (10%) between genotypes as the rest of the genome (11-13%) (Smith *at al.* 1997a). This difference is surprising given that both viruses are believed to replicate using a highly structured IRES, located in the 5'UTR. The IRES would be expected to be under the same structural constraints as base paired synonymous nucleotides within the ORF. There is no obvious reason why the 5'UTR of HCV should be under greater evolutionary constraint than that of HGV/GBV-C. Consequently, as the short term rates of sequence change for the two viruses have been shown to be comparable, the greater divergence of the HGV/GBV-C 5'UTR is

further evidence of an earlier time of origin for HGV/GBV-C than the major genotypes of HCV (Simmonds and Smith 1999a).

1.10 QUESTIONS TO BE ANSWERED

A number of previous studies have shown that constraints on sequence variation exist within the polyprotein coding region of HCV and to an even greater degree HGV/GBV-C (Smith and Simmonds 1997; Simmonds and Smith 1999b) In particular, restrictions on variation at synonymous codon sites and clustering of covariant substitutions have been documented within the ORFs of both HCV and HGV/GBV-C. It has been suggested that such biases are due to the presence of functionally important and evolutionarily conserved RNA structure. Although, such structure has been demonstrated within the 5' and 3' UTRs of both viruses, little progress has been made toward the identification and characterisation of structures residing within the coding regions of RNA viruses.

In order to fully define the existence, position and confirmation of RNA structure within the coding regions of HCV, HGV/GBV-C and related viruses their genomes were analysed with a combination of thermodynamic prediction of folding free energies (FFEs), evolutionary conservation of minimum energy structures between virus genotypes, suppression of synonymous variability and analysis of covariant and semi covariant substitutions in thermodynamically favoured structures (Chapters 3 and 4). Each of the predictive methods has provided evidence for conserved RNA

secondary structures in the core gene and NS5B coding region of HCV and throughout the entire genome of GBV-C (Cuceanu *at al.* 2001; Tuplin *at al.* 2002).

The existence of predicted RNA structures in the core gene and NS5B coding regions of HCV were determined using a process of controlled nuclease mapping, in which RNA transcripts were analysed under conditions which maintain potential long-range RNA interactions (Chapter 6). Electron microscopy was used to directly visualise the RNA folding structure of HGV/GBV-C RNA transcripts including the RNA folding conformation of the complete virus genome (chapter 5).

Whilst the role of the RNA structures identified in this study is not currently understood they are providing a starting point for functional studies using the HCV replicon.

CHAPTER 2

MATERIAL AND METHODS

2. MATERIALS AND METHODS

2.1 POLYMERASE CHAIN REACTION

2.1.1 EXTRACTION OF VIRAL RNA FROM SERUM SAMPLES

Viral RNA was extracted through incubating 100 µl of serum sample with 400 µl of lysis buffer (50 mM Tris-HCl pH 8.0, 0.1 M NaCl, 0.5% SDS, 1 mM EDTA, 40 µg/ml polyadenylic acid, 1 mM proteinase K) at 37 °C for 90 min. 450 µl of phenol (Rathburn Chemicals Ltd.), was then added to each extraction tube and vortexed thoroughly. After centrifugation at 12,000 g for 10 min the aqueous layer was transferred to a new Eppendorf tube containing 450 µl of chloroform: isoamylalcohol (50:1). This mixture was again vortexed thoroughly and centrifuged at 12,000 g for 10 min. The aqueous layer was transferred to a new Eppendorf tube. The nucleic acid was precipitated overnight at -20°C, after the addition of 2.5 volumes of ice-cold 100% ethanol and 0.1 volumes of sodium acetate (pH 5.2). The nucleic acid was collected by centrifugation at 12,000 g at 0°C for 15 min. The pellet was then washed in 80% ethanol and air dried at 42° for approximately 15 min. The dried pellet was resuspended in 30 µl of nuclease free water (Ambion). During the process known samples of positive and negative serum samples were extracted in parallel in order to act as negative and positive controls to ensure that no contamination was present.

2.1.2 REVERSE TRANSCRIPTION OF VIRAL RNA

Complementary DNA (cDNA) can be synthesised from single-stranded RNA using a reverse transcriptase reaction. Synthesis of cDNA was carried out using 5 µl of extracted RNA, 10 Units of AMV reverse transcriptase (Promega), 40 Units of RNasin Ribonuclease inhibitor (Promega), 1 mM each of dGTP, dATP, dTTP and dCTP (Promega) and 0.5 µM of the antisense primer (also used in corresponding PCR reaction) in a reaction buffer containing 50 mM Tris-HCl, pH 8.3, 50 mM KCl, 10 mM $MgCl_2$, 10 mM DTT and 2.5 mM spermidine. Each sample was covered with a drop of liquid paraffin oil and incubated on a hot block at 48°C for 30 min followed by 98°C for 2 min in order to inactivate the reverse transcriptase.

2.1.3 PCR AMPLIFICATION

The polymerase chain reaction (PCR) is a technique for the *in vitro* amplification of specific DNA sequences by the simultaneous primer extension of complementary strands of DNA. The sensitivity of PCR can be reduced when using low copy number DNA template in which case nested, or hemi-nested, PCR can be used. In this method the PCR is repeated using a pair of primers internal to the original primers, or in the case of hemi-nested PCR, one new internal primer and the opposing primer from the first PCR reaction. Hemi nested PCR was used to amplify HGV/GBV-C 3'UTR and NS5B regions from the reverse transcribed cDNA product of human serum (section 2.12). Single round PCR amplification was used to amplify

subgenomic regions of both HGV/GBV-C and HCV from purified cloned cDNA as the template copy number was higher (section 2.2.6). Internal primers were used for single round PCR amplification of HGV/GBV-C and outer primers for HCV (Table 2.1).

The primary or single round PCR reaction was carried out in a 50 µl volume containing 5 µl of cDNA template (section 2.1.2), 0.5 U of DNA Taq polymerase (Promega), 0.3 mM each of dGTP, dATP, dTTP and dCTP and 0.25 µM of each outer primer (sense and antisense primers) in PCR reaction buffer containing 50 mM KCl, 10 mM Tris-HCl (pH 9.0), 0.1% Triton X-100 and 1.5 mM $MgCl_2$. Each sample was covered with a drop of liquid paraffin oil and transferred to a Techne Genius thermal cycler.

The secondary PCR reaction was carried out in a 20 µl volume, using 1 µl of the primary PCR reaction as the DNA template. 0.2 µM of each inner primer was used and the reaction set up as previously for the primary PCR conditions.

Known positive and negative controls were included to check the integrity of the reaction and to ensure that there was no contamination. A complete list of primer sequences and annealing positions for each subgenomic region amplified is shown in Table 2.1.

The template denaturation, primer annealing and strand elongation conditions for the PCR amplification reaction were as follows. 30 cycles were performed in each case:

HGV/GBV-C NS5B region and 3'UTR (first round): 94°C for 18 s; 58 °C for 21 s: 72 °C for 90 s

HGV/GBV-C NS5B region (second and single round): 94°C for 18 s: 58 °C for 21 s: 72 °C for 90 s

HGV/GBV-C 3'UTR (second and single round): 94°C for 18 s: 55 °C for 21 s: 72 °C for 90 s

HCV core gene and NS5B region (single round): 94°C for 18 s: 55 °C for 21 s: 72 °C for 120 s

2.1.4 ANALYSIS OF PCR PRODUCT

The amplified DNA was visualised under UV light after electrophoresis through a 1% agarose gel in 1×TAE buffer stained with 0.07 μg/ml ethidium bromide (EtBr). The gel was run for 30 min at 150V. Amplified DNA was detected due to the presence of ethidium bromide, which intercalates into the DNA and exhibits fluorescence under UV light.

2.2 CLONING OF PCR PRODUCT

2.2.1 USING pGEM-T EASY VECTOR

A number of thermostable polymerases, such as Taq DNA polymerase, add a single deoxyadenosine (A), in a template independent manner, to the 3'-ends of PCR amplified fragments. The pGEM-T EASY vector (Promega) is supplied pre-cut with

Table 2.1 Primers used for PCR amplification of HCV and HGV/GBV-C.

Region	Name	Position	Orientation	Sequence
HGV/GBV-C NS5B	AT1	8411-8429	I & O/S	CCATCACACGGTGGGTCATCAT
	AT4	8875-8894	I/AS	ACGATGAGCAGGGCTAAGAA
	AT5	8906-8927	O/AS	CAACAGATGAATTTAGTTCACC
HGV/GBV-C 3'UTR	Z3580#	8827-8836	O/S	GGTGGTNCATCAATTGGATT
	Z3281#	8873-8894	I/S	GGTTCTTAGCCCTGCTCATC
	Z3282#	9212-9231	O & I/AS	AGTAGAACCCGGCCTTTGGG
HCV CORE	HCV_C_1*	-23 - -2	S	TAATACGACTCACTATAGGGTAATACGACTCACTATAGGG
	HCV_C_2	157-176	A	CGAGGTTGCGACCGCTCGGA
	HCV_C_3	327-456	A	ACCTAYGCGGGGTCWK
	HCV_C_4	497-516	A	GCAACCAGGAAGGTTCCCTG

Region	Name	Position	Orientation	Sequence
HCV NS5B	HCV1A_N_1*ψ	8717-8736	S	TAATACGACTCACTATAGGGAGGCCCTCGATTGCGAGATC
	HCV2A_N_1*ǂ	8967-8987	S	TAATACGACTCACTATAGGGAGAACCTCAACTTTGAGATG
	HCV_N_2	9066-9086	A	TYACYGCCCARTTGAAGAGR
	HCV_N_3	9067-9086	A	ACGCTGTGATAAATGTCTCC
	HCV1A_N_4ψ	9167-9186	A	TGTTTACCCCAACCTTCATC
	HCV2A_N_4ǂ	9167-9186	A	CTAATGTGTGCCGCTCTACC
pGEM-T easy vector	M13 FOR[$]	2959-2976α	S	GTTTTCCCAGTCACGAC
	M13 REV[$]	177-192	A	CAGGAAACAGCTATGAC

*** T7 phage polymerase promoter included. Ψ HCV genotype 1a specific. ǂ HCV genotype 2a specific. # Primers by permission of the Scottish National Blood Transfusion Service. $ Promega corporation.**

EcoR V restriction enzyme and with the addition of deoxythymidine (T) to the 3' terminus at both ends. This reduces the likelihood of vector re-ligation and provides a compatible overhang for the terminal A produced by Taq polymerase.

The pGEM-T EASY vector contains the promoter and a small region of the β-galactosidase (β-*gal*) gene, known as the α-domain, into which the multiple cloning site (MCS) is inserted. The competent cells produce β-gal peptide, which is truncated at the N-terminus. The truncated peptide combines with the α-domain to produce a functional β-gal protein. The cells are supplied with 5-bromo-4-chloro-3-indolyl-β-D-galactsidase (X-Gal) (plus an inducer of the enzyme such as isopropyl-thiogalactoside (IPTG)) which, as a substrate of the β-gal protein, is hydrolysed to produce a precipitate that colours the colonies blue. When an insert is ligated into the MCS of the vector the β-*gal* sequence is disrupted. This prevents the formation of the functional peptide and white colonies are normally produced.

2.2.2 PREPARATION OF PCR PRODUCT FOR CLONING

The amplified DNA band was excised from the 1 % (w/v) agarose gel after electrophoresis and purified using a QIAquick Gel Extraction Kit (QIAGEN), following the manufacturers instructions. After excision the gel fragment was weighed in a 1.5 ml microcentrifuge tube. 300 µl of solubilisation buffer QG were added for every 100 mg of gel slice in order to dissolve the gel fragment and to achieve an optimum pH for binding DNA to the silica-gel membrane. This was then mixed vigorously by vortexing and the reaction incubated at 50°C for 10 min, or until

the gel slice had completely dissolved. The DNA was then precipitated from solution by the addition of 100 µl of isopropan-2-ol for every 100 mg of gel slice. This was then added to a QIAquick spin column in a 2 ml collection tube and the DNA bound to the silica-gel membrane by centrifugation at 12,000 g for 60 sec. The flow through was discarded and any remaining traces of agarose were removed by the addition of 500 µl of solubilisation buffer QG followed by centrifugation at 12,000 g for 60 sec. The flow through was again discarded and excess salts removed by the addition of 0.75 ml of ethanol containing buffer PE, followed by centrifugation at 12,000 g for 60 sec. The flow through was discarded and a final 60 sec centrifugation at 12,000 g removed any residual buffer. The column was placed in a clean 1.5 ml microcentrifuge tube and 30 µl of nuclease free water was added to the centre of the QIAquick spin column. After a 1 min incubation the DNA was eluted by centrifugation at 12,000 g for 60 sec.

2.2.3 LIGATION OF INSERT USING pGEM-T EASY VECTOR

Each ligation was setup in a 0.75 ml microcentrifuge tube with 2µl of purified HGV/GBV-C NS5B or 3'UTR PCR product (section 2.1.3), 5 µl of 2× rapid ligation buffer (Promega), 1 µl of pGEM-T Easy vector (1µg/µl) (Promega), 1 µl of T4 DNA ligase (3U/µl) (Promega) and made up to 10 µl with nuclease free water. This was mixed by pipetting and incubated overnight at 4°C. A ligation control was setup substituting the PCR product for 2 µl of control DNA (4ng/µl) supplied with the kit. A background control was also set up, to test the integrity of the pGEM-T Easy

vector T overhangs, in which the ligation was carried out with no insert DNA but the total volume made up with nuclease free water.

2.2.4 TRANSFORMATION REACTION

JM109 competent cells (Promega) with the following genotype were used for the transformation: *end*A1, *rec*A1, *gyr*A96, *thi*, *hsd*R17 (r_k^-,m_k+), *rel*A1, *sup*E44, Δ(*lac-pro*AB), [F', *tra*D36, *pro*AB, laq^{Iq}ZΔM15].

The ligation mix was centrifuged briefly to collect the reaction mix at the bottom of the tube, 2 µl of which was transferred to a 1.5 ml microcentrifuge tube on ice. Competent cells were thawed on ice for about 5 min and mixed by gently flicking the tube. 50 µl of the competent cells were transferred to the aliquoted ligation reaction, mixed by flicking the tube and incubated on ice for 20 min. The tube was then heat shocked for 45 s at 42°C and returned to the ice for 2 min. 950 µl of SOC medium was then added and the tube incubated at 37°C for 1.5 h in an orbital incubator at 12,000 g.

Luria Broth (LB) agar plates, containing 100 µg/ml ampicillin, 0.5 mM (IPTG) and 80 µg/ml X-Gal were prepared. Ampicillin (Sigma) was used for selection to ensure the presence only of those transformants with the plasmid containing the ampicillin resistance gene (β-*lactamase*).

After incubation 100 µl of the transformed cells were plated out on the LB agar selective plates and incubated overnight at 37°C.

2.2.5 SCREENING OF TRANSFORMANTS FOR DNA INSERTS

Successful cloning of an insert into the pGEM-T Easy vector leads to the disruption of the *β-gal* gene, which usually results in white colonies on an LB selective plate. White colonies were screened for the presence and orientation of HGV/GBV-C NS5B and 3'UTR cDNA inserts by single round PCR amplification (section 2.1.3).

A sterile toothpick was used to transfer the edge of a white colony into a 0.5 ml PCR tube containing 20 µl of PCR mix (0.3 U of DNA Taq polymerase (Promega), 0.3 mM each of dGTP, dATP, dTTP, and dCTP and 0.25 µM of each primer (sense and antisense primers). In order to determine the insert orientation primers were designed to amplify across the MCS boundary; with one designed to hybridize to the vector and the other to the cDNA insert itself (primer sequences and annealing positions are shown in table 2.1). PCR amplification was performed across both the T7 (5') and SP6 (3') boundary of the pGEM-T Easy vector construct so that each colony was checked twice, for both positive and negative PCR amplification. Each solution was covered with a drop of liquid paraffin oil and placed on a Techne Genius thermal cycler for 30 cycles of 94°C for 30 s, 50°C for 90 s and 72°C for 120 s. The samples were then analysed by electrophoresis through a 1% agarose/TAE gel (w/v) containing 0.07 µg/ml EtBr next to a molecular weight marker for reference (2.1.4). Recombinant transformants of the appropriate size and orientation were identified and the corresponding colonies on the agar plate selected.

2.2.6 PREPARATION OF PLASMID DNA

A sterile toothpick was dabbed on the edge of selected recombinant transformant colonies. This was then used to inoculate 3 ml of LB broth containing 100 μg/ml ampicillin. The inoculated culture media was incubated at 37 °C overnight in an orbital incubator at 150 g.

The plasmid DNA was collected and purified from the overnight culture using a QIAprep miniprep (QIAGEN) following the manufacturers instructions and using the buffers provided. The bacterial cells were harvested from the overnight culture by centrifugation at 9,000 g for 5 min, after which the pelleted cells were retained and the supernatant discarded. The bacterial pellet was resuspended in 250 μl of buffer P1, containing RNase A. The cells were lysed using an alkaline lysis method by the addition of 250 μl of buffer P2 following which the tube was inverted a number of times. After 5 min the lysis reaction was terminated by the addition of 350 μl of buffer N3, this buffer also increases the salt concentration of the suspension, optimising the conditions for DNA adsorption to silica gel within a spin column. The lysate was then cleared by centrifugation at 12,000 g for 10 min and the supernatant applied to a QIAprep spin column. The plasmid DNA was bound to the silica gel by centrifugation at 12,000 g for 1 min and the flow through discarded. The bound plasmid DNA was washed by the addition of 0.5 ml of buffer PB followed by centrifugation at 12,000 g for 1 min, which removed any endonucleases that may have been released during cell lysis. After the flow though was discarded, salt was removed from the plasmid DNA by the addition of 0.75 ml of ethanol containing buffer PE followed by centrifugation at 12,000 g for 1 min. The flow

through was discarded and the columns were again by centrifugation at 12,000 g for 1 min to remove any residual wash buffer. The spin column was placed in a clean 1.5 ml collection tube and 50 µl of nuclease free water was added to the centre of the QIAquick gel. After 1 min incubation the plasmid DNA was eluted by centrifugation at 12,000 g for 1 min.

2.2.7 DNA SEQUENCING

The nucleotide sequence of each cDNA insert within the purified plasmid DNA constructs was determined by cycle sequencing, which also further confirmed the orientation of the cDNA insert. Sequencing was carried out in the presence of limiting radiolabeled dideoxynucleoside triphosphates (ddNTPs) which terminate synthesis at specific bases. Sequencing was performed using the "Thermo Sequenase™ Radiolabeled Terminator Cycle Sequencing Kit" supplied by United States Biochemical (USB). Thermo Sequenase DNA polymerase was used for strand extension. This enzyme was created by in vitro genetic manipulation and is a variant of bacteriophage T7 DNA polymerase, without 3' or 5' exonuclease activity. The enzyme is particularly suited to this activity as it is able to efficiently incorporate strand terminating ddNTPs.

The reaction was performed directly using cloned cDNA as template and involved the incorporation of one of four terminating [α-33P]ddNTPs into the sequencing reaction products at the 3'end. In the case of PCR product, the template was pre-treated with combination of Exonuclease I (ExoI) and Shrimp Alkaline Phosphatase

(SAP) to eliminate any primer or dNTPs which were not incorporated into the PCR product (both enzymes were supplied by USB). Pre treatment of PCR amplification product was carried out in a 10 µl volume containing 5 µl of PCR product (approximately 100 ng/µl), 2 U of SAP, 10 U of ExoI and made up to 10 µl with nuclease free water (Ambion). The reaction mix was incubated at 37°C for 15 min before denaturing at 80°C for a further 15 min.

A termination mix was prepared for each dNTP in separate 0.5 ml eppendorf tubes each containing 1 µl of nucleotide master (7.5 µm each of dATP, dCTP dGTP and dTTP), 0.5 µl of the appropriate [α-33P]ddNTP (0.45 µCi/µl) (Amersham) and made up to 2.5 µl with nuclease free water (Ambion). A reaction mixture was set up containing 2 µl of reaction buffer (260 mM Tris-HCl, pH 9.5 65 mM $MgCl_2$), 0.15 mM of the appropriate primer, 4 U of Thermo Sequenase polymerase and made up to 20 µl with nuclease free water. The cycle termination reactions were then prepared in four separate 0.5 µl eppendorf tubes (one for each [α-33P]ddNTP), each containing between 100 – 200 ng of DNA template, 4 µl of reaction mixture, 2.5 µl of termination mix and made up to 10 µl with nuclease free water. Each termination reaction was covered with a drop of liquid paraffin and transferred to a Techne Genius thermal cycler for 30 cycles of 95°C for 30 sec, 55°C for 30 sec and 72°C for 30. 6 µl of each termination reaction was transferred to a fresh 0.5 ml Eppendorf tube containing 4 µl of denaturing stop solution (95 % formamide, 20 mM EDTA, 0.05 bromophenol blue, and 0.05% xylene cyanol).

2.2.8 ANALYSIS OF SEQUENCING PRODUCT

Sequencing products were analysed by electrophoresis through a 5% denaturing acrylamide gel. The quality of electrophoresis gel is an important factor which limits the extent of sequence information that can be determined from a sequencing experiment. Consequently, electrophoresis grade reagents were used and glass plates cleaned by swabbing with ethanol, methanol and acetone prior to pouring the gel. The gel contained the following reagents: 21 g urea (Ana-BDH), 5% v/v Ultrapure Sequagel concentrate (acryl:bis-acryl = 19:1) (National Diagnostics) 5 ml of 10× Sanger TBE, 0.05 g ammonium persulphate and distilled water up to 50 ml. When all the urea had dissolved 20 µl of TEMED (N,N,N,N-tetramethylethlediamine) (Sigma) was added and the gel solution was poured into the plate assembly.

Gels were prepared at least 2 hours prior to use and were pre-run for 15 min in 1× Sanger TBE electrophoresis buffer before loading the samples. Prior to loading samples were denatured by heating at 95 °C for 5 min. Gels were run for between 1 and 3 hours, depending on the fragment size of interest. Gels were dried and exposed overnight on BIOMAX autoradiography film (Eastman Kodak).

2.3 TRANSCRIPTION IN VITRO

2.3.1 LINEARISATION OF PLASMID TEMPLATE

T7 and SP6 DNA dependent RNA polymerases were originally isolated from bacteriophage infection of their respective *Escherichia coli* (E. coli) hosts. They are both template dependent polymerases with distinct and highly specific promoter requirements. Before cloned DNA can be used as a template for transcription *in vitro* it must be linearised, by restriction digest downstream of the insert DNA, to prevent the RNA polymerase reading through the insert into the vector sequence.

10 µg of purified plasmid DNA template containing either HGV/GBV-C NS5B region or 3'UTR cDNA inserts were incubated with 5 U of Spe I restriction enzyme (Promega), 2 µl of 10 × reaction buffer B (Promega) and made up to 20 µl with nuclease free water. The reaction was mixed by pipetting and incubated for 3 hrs at 37°C

2.3.2 PREPARATION OF LINEAR PLASMID FOR TRANSCRIPTION *IN VITRO*

The plasmid digestion mix was made up to 200 µl with DEPC-water. 200 µl of phenol (Sigma) was added and the tube vortexed. After centrifugation at 12,000 g for 5 min the aqueous layer was transferred to a new microcentrifuge tube, containing 200 µl of chloroform, which was again centrifuged at 12,000 g for 5 min.

The aqueous layer was then transferred to a new microcentrifuge tube and precipitated at -20°C after the addition of 2.5 volumes of 100% ethanol and 0.1 volumes of sodium acetate (3.5M, pH 8). The DNA was collected by centrifugation at 12,000 g for 15 min. The supernatant was removed and the pellet washed by the addition of 500 µl of 70% ethanol (nuclease free) followed by centrifugation at 12,000 g for 15 min. The supernatant was removed and the pellet dried for 5 min at 90 and then resuspended in 20 µl of nuclease free water.

2.3.3 *IN VITRO* TRANSCRIPTION

The transcription reaction was performed with a "T7 MEGAscript high yield transcription kit" (Ambion), according to the manufacturers instructions. Firstly the 10 × reaction buffer and four ribonucleotides (ATP, CTP, GTP and UTP) were vortexed and briefly centrifuged. The ribonucleotides were then stored on ice along with the T7 RNA polymerase enzyme. The reaction was then set up in an RNase free 0.75 ml microcentrifuge tube containing (7.5 mM) of each ribonucleotide either 1 µg of linear cloned cDNA or 0.1-0.2 µg of T7 annealed PCR product, 2 µl of 10× reaction buffer (Ambion), 2 µl of T7 RNA polymerase enzyme mix (Ambion) and made up to 20 µl with nuclease free water (Ambion). The reaction was then mixed by gently flicking the tube, centrifuged briefly and incubated overnight at 37°C.

2.3.4 PURIFICATION OF RNA

The transcribed RNA was treated with 2 U of DNase I (Ambion) for 30 minutes at 37°C in order to remove the DNA template and then made up to 200 µl with 15 µl of 5 M ammonium acetate and 164 µl of nuclease free water (Ambion). 200 µl of low pH phenol (pH 4.5) was then added and the tube vortexed. After centrifugation at 12,000 g for 5 min the aqueous layer was transferred to a new microcentrifuge tube, containing 200 µl of chloroform, which was again vortexed and centrifuged at 12,000 g for 5 min. The aqueous layer was transferred to a new microcentrufuge tube and the RNA precipitated for 15 min at –20 °C after the addition of 1 equal volume of isopropan-2-ol. The RNA was then collected by centrifugation at 12,000 g for 15 min and the pellet dried for 5 min at 90 °C. The dried pellet was resuspended in 20 µl of nuclease free water with 1µl of Anti-RNAse (Ambion).

2.3.5 ANALYSIS OF TRANSCRIPTION PRODUCT

1 µl of purified RNA was analysed by electrophoresis through a 5% acrylamide (acryl: bis-acryl = 19:1), denaturing (7 M urea) gel in 1×TAE buffer against an RNA molecular weight marker (Century Marker Plus or Millennium Marker) (Ambion) in order to check the size and integrity of the transcription products. 1 µl of purified transcription product was made up to 20 µl with 14 µl of nuclease free water and 5 µl of loading buffer (95% formamide, 0.025% xylene cyanol, 0.025% bromophenol blue, 18 mM EDTA and 0.025% SDS) (Ambion). This was then heated to 95°C for

2 min, in order to denature the RNA, and run on a 5% denaturing acrylamide gel for 60 min at 150 V.

The RNA was visualised by staining the gel in a tank for 1 hour under a solution of methylene blue (0.04 % methylene blue (w/v), 0.5 M sodium acetate) and then rinsing with distilled water. Alternatively the integrity of 1 µl of purified RNA was assessed by visualisation under UV light after electrophoresis through a 1% (w/v) agarose gel in 1×TAE buffer with 0.07 µg/ml ethidium bromide.

2.4 TRANSMISSION ELECTRON MICROSCOPY

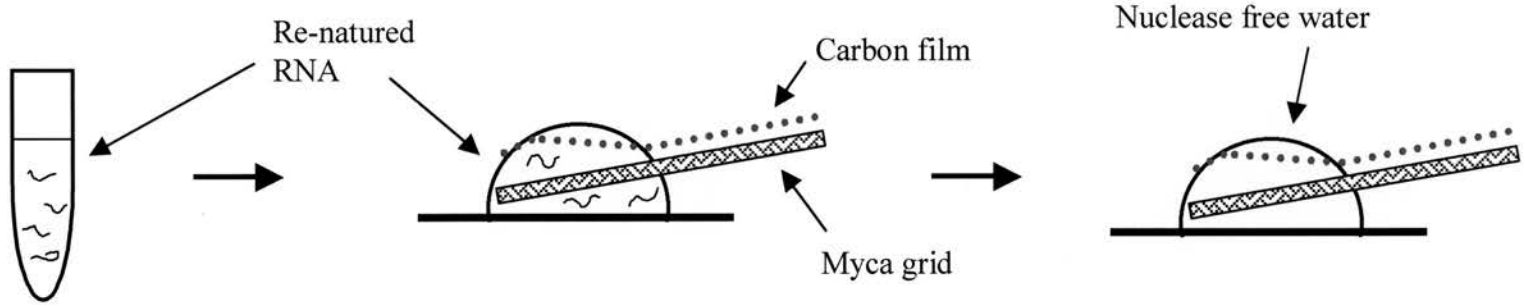
2.4.2 RNA ADSORPTION AND ELECTRON MICROSCOPY

Purified HGV/GBV-C NS5B, 3'UTR or complete genome transcript RNA was diluted to 3 µg/ml in a volume of 150 µl with nuclease free water. $MgCl_2$ and NaCl were added to a final concentration of 10 mM to stabilise RNA secondary structures. Approximately 1.2×10^{13} molecules of biotinylated oligonucleotides were hybridised to specific stem loop structures (Table 1.2). The mix was then heated to 70 °C for 5 min and cooled slowly to room temperature for approximately 2 hours to allow RNA structures to reform and overlain with 10 nm streptavidin-coated gold particles. RNA was then adsorbed onto a carbon film and negatively stained with a solution of uranyl acetate (2 %, w/v, pH 4.5), (Agar), (Fig. 2.1). 200 µl drops of diluted RNA, nuclease free water and uranyl acetate were placed on individual segments of parafilm. A 3 mm square segment of carbon covered mica was dipped into the RNA

Table 2.2 Biotinylated oligonucleotides used for hybridisation to RNA structures in HGV/GBV-C.

Region	Stem Loop	Positions	sequence
5'UTR	Domain IVa	326-393	GAGAGACATTGAAC
NS5B	$SL_{NS5B}III$	8710-8723	GCCAACTCCGCCCA
NS5B	$SL_{NS5B}IV$	8560-8573	GCTGCCGGTCCGTG
3'UTR	3'SLII	9176-9190	CCCCTCAAATCACA

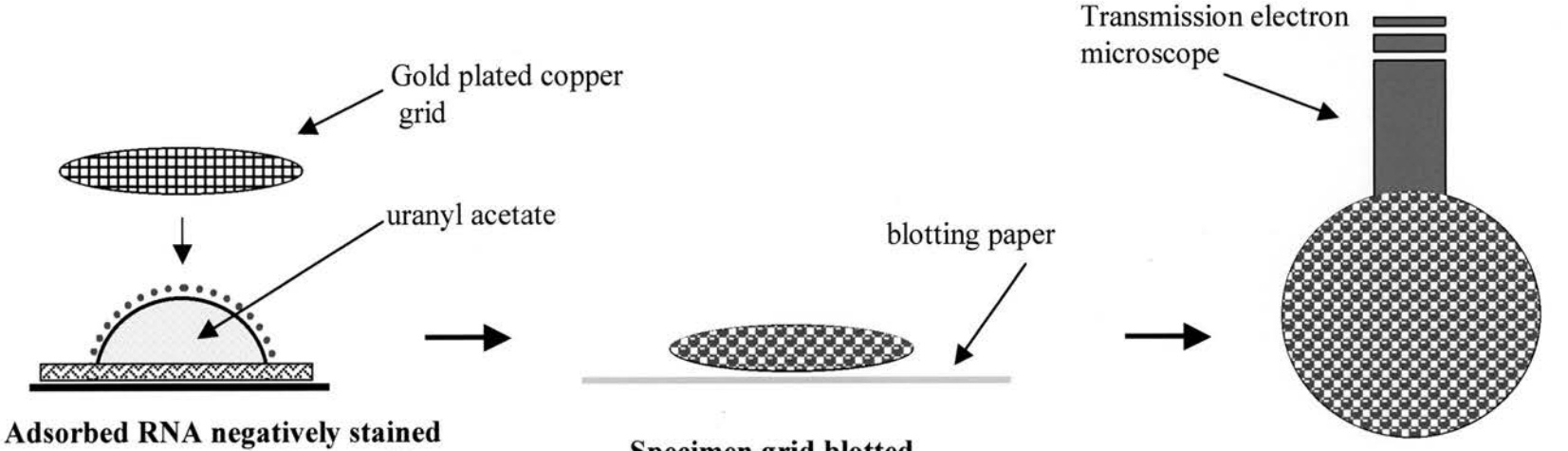
Figure 2.1 Electron microscopy of *in vitro* transcribed viral RNA



RNA heated and re-natured in physiological cation conditions

RNA adsorbed into carbon film for 60 sec

Carbon film washed for 10 sec in nuclease free water



Adsorbed RNA negatively stained with uranyl acetate. Carbon film released from myca and mounted on grid

Specimen grid blotted dry

Micrograph recorded on a transmission electron microscope

droplet for 60 sec, to allow the adsorption of RNA onto the carbon film. The carbon covered mica was then washed in the nuclease free water for 10 sec and negatively stained for 90 sec in the uranyl acetate droplet. The carbon film was then mounted on a gold plated copper mesh grid (300 mesh) (Agar) and dried on blotting paper for 1 hour. Micrographs were recorded on Agfa Scientia 23D56 electron image film electron image sheet films at calibrated magnification on a Philips CM10 transmission electron microscope operated at 100 kV.

2.5. NUCLEASE MAPPING OF RNA SECONDARY STRUCTURE

2.5.1 PARTIAL NUCLEASE DIGESTION OF RNA

4μg of HCV core gene or NS5B RNA transcript was made up to 53.4 μl with nuclease free water and 6 μl of RNA structural buffer (10 mM Tris pH 7, 100 mM KCl, 10 mM) (RNA structural analysis kit, Ambion), melted at 70 °C for 5 min and then cooled at room temperature for 45 min. After the addition of 6 μg of yeast carrier RNA (10 μg/μl), (RNA structural analysis kit, Ambion) 9μl of the reaction mix was aliquoted into six individual 0.75 ml microcentrifuge tubes for nuclease digestion. 1 μl of nuclease free water was added to tube 1, 1 μl of RNA nuclease T_1 (RNA structural analysis kit, Ambion) was added to tubes 2 and 3 at 0.02 U/μl and 0.14 U/μl respectively. The same procedure was followed for tubes 4 to 6 but using RNA nuclease V_1 (RNA structural analysis kit, Ambion) at 0.005 U/μl and 0.002 U/μl. The digestion was run at room temperature for 15 min before being

stopped by the addition of 20 µl of inactivation/precipitation solution (RNA structural analysis kit, Ambion) followed by alcohol precipitation as described in section 2.3.4.

2.5.2 PRIMER EXTENSION BY REVERSE TRANSCRIPTION

1 µg of resuspended RNA was made up to 10 µl with 1 µg of one of three reverse primers for each subgenomic region (table 2.1) and nuclease free water. The RNA was then heated to 70 °C for 5 min and cooled quickly on ice to minimise the formation of structure and allow the hybridisation of primer to template. The reverse transcription reaction was then set up by the addition of 12.5 mM each of dGTP, dTTC and dCTC (Promega), limiting [α-33P]ATP (10 µCi/µl) (Amersham), 5 µl of reaction buffer (50 mM Tris-HCl (pH 8.3 at 25°C), 75 mM KCl, 3 mM $MgCl_2$, 5 mM DTT) (Promega), 25 U RNasin (Promega), 200 U of M-MLV reverse transcriptase (RNase H deficient) (Promega) and made up to 25 µl with nuclease free water. Each sample was then covered with a drop of paraffin, incubated at 42 °C for 60 min and chased after 20 min by the addition of 12.5 mM of dATP.

2.5.3 PREPERATION OF RADIOLABLED cDNA FOR ANALYSIS

In order to remove excess salt, which would decrease band resolution when the cDNA was analysed by electrophoresis, the radiolabeled cDNA product was purified using a QIAquick PCR Purification kit (QIAGEN) following the manufacturers instructions and using the buffers provided.

Five volumes of buffer PB were added to 1 volume of cDNA product in order to achieve an optimum DNA binding pH for silica-gel binding. This was then added to a QIAquick spin column in a 2 ml collection tube and the cDNA bound to the silica gel by centrifugation at 12,000 g for 30 sec. The flow through was discarded and excess salt removed from the cDNA by the addition of 0.75 ml of ethanol containing buffer PE, followed by centrifugation at 12,000 g for 30 sec. The flow through was discarded and a final centrifugation at 12,000 g for 30 sec removed any residual buffer. The column was placed in a clean 1.5 ml microcentrifuge tube and 30 µl of nuclease free water was added to the centre of the QIAquick gel. After a 1 min incubation the cDNA was eluted by centrifugation at 12,000 g for 60 sec.

2.5.3 ANALYSIS OF RADIOLABELED cDNA

Purified radiolabeled cDNA was analysed by electrophoresis through a 5 % acrylamide gel (w/v) (acryl: bis-acryl = 19:1), denaturing (7 M urea) gel in 1 × Sanger TBE buffer against radiolabeled cycle sequencing products of the same DNA

sequence in order to map primer extension termination products to defined genome regions.

4 µl of cDNA product was made up to 7 µl with denaturing loading buffer (95 % formamide (v/v), 0.05% xylene cyanol (v/v), 0.05 % bromophenol blue (w/v), 20 mM EDTA (w/v)). This was then heated to 95 °C for 5 min, in order to denature the cDNA and analysed on a 5 % (w/v) acrylamide gel for between 2 hrs 30min and 60 min depending on the length of the cDNA to be analysed. Gels were dried and exposed overnight on BIOMAX autoradiography film (Eastman Kodak). (see section 2.2.8 for full details of this procedure)

2.6. THERMODYNAMIC PREDICTION OF FOLDING FREE ENERGY (FFE) AND RNA STRUCTURE PREDICTION

Prior to analysis all sequences were aligned by hand using the SIMMONIC sequence editor package (Simmonds and Smith 1999). All FFE calculations were determined using MFOLD 3.1 (http://.bioinfo.rpi.edu/~zukerm/), using the implementation available in the program ZIPFOLD on default setting which allow FFEs of RNA sequences to be rapidly determined (Zuker 2003). MFOLD and ZIPFOLD provide computational speed and accuracy of structural prediction independently of comparative sequence information required by phylogenetic structure prediction methods. In order to calculate folding free energy differences (FFED) the FFE of each native sequence was compared to those of 50 independent sequence randomisations of the original sequence.

All sequence randomisation methods were carried out using the SIMMONIC sequence editor package. A number of sequence randomisation algorithms were implemented including those which preserve dinucleotide frequencies and local biases (CDR and CDS), three codon orientated randomisation methods (CLR, CLS and COS) and a standard nucleotide randomisation method. The GenBank Accession numbers of all sequences analysed is shown in appendix.

Specific structural predictions of RNA secondary structure were made for those regions of the polyprotein coding regions showing suppression of synonymous substitutions, clustering of covariant substitutions and excess FFEDs from sequence order randomised controls. In practice this meant the core gene of HCV and the NS5B regions of both HCV and HGV/GBV-C; predictions were also made for the 3'UTR region of HGV/GBV-C. Conservation of each structure was assessed by parallel folding of each region in all available epidemiologically unlinked complete genome sequences and comparison of covariant and semi-covariant substitutions. Alignments and a complete list of GenBank Accession numbers are shown in the appendix.

CHAPTER 3

FOLDING FREE ENERGY DIFFERENCES

3.1 INTRODUCTION

A number of previous investigations have attempted to reconstruct times of divergence between genotypes of both HCV (Ogata et al., 1991; Okamoto et al., 1992; Abe et al., 1992; Smith et al., 1997b), and HGV/GBV-C (Nakao et al., 1997; Khudyakov et al., 1997), based on estimated rates of sequence change within chronically infected individuals over time or between divergent virus genotypes. Such studies have revealed that a number of constraints on sequence change appear to act on the genomes of both viruses; limiting the use of such a "molecular clock" based approach in calculating rates of divergence over time (more detail in section 1.9) (Smith and Simmonds, 1997a; Tuplin et al., 2002).

In HCV it was shown that synonymous variability, although generally presumed to be selectively neutral, was suppressed within subgenomic regions of the virus genome (Ina et al., 1994; Smith and Simmonds, 1997a; Tuplin et al., 2002). The greatest constraints on synonymous variability were observed across the core gene and NS5B region; within the 5' and 3' coding regions of the virus respectively (Smith and Simmonds, 1997a; Tuplin et al., 2002). Even greater constraints on sequence divergence have been observed in HGV/GBV-C (Simmonds and Smith, 1999b; Cuceanu et al., 2001). In particular, a large disparity between the short- and long-term rates of sequence change has been noted (Simmonds and Smith, 1999b). In contrast to HCV, constraints on variability at synonymous sites have been observed along the length of the virus genome and were more extreme. In further analysis, of divergent genotypes, it was noted that covariant substitutions clustered within regions of suppressed synonymous variability within the genomes of both

viruses, yet were again more numerous in HGV/GBV-C than HCV (Smith and Simmonds, 1997a; Simmonds and Smith, 1999b; Cuceanu et al., 2001).

It has been suggested that these observed constraints on sequence change may result from the presence of evolutionarily conserved RNA structure (Smith and Simmonds, 1997a; Simmonds and Smith, 1999b). In a folded genome, sequence change in regions that are internally base paired would require simultaneous nucleotide substitutions on either side of stem loop structure to maintain base pairing. It would also lead to a much lower rate of sequence change than in unpaired regions, as sites in which substitutions would not disrupt RNA structure will quickly become saturated. Whilst these results are consistent with the presence of conserved RNA structure within the genomes of both viruses, the more extreme biases observed in HGV/GBV-C are suggestive of more extensive structure than within the genome of HCV.

In the present study, in order to investigate the extent of sequence dependent RNA structure within the genomes of HCV and HGV/GBV-C, folding free energy (FFE) was determined across the virus genomes and was compared to that of 50 independently generated sequence order controls for corresponding regions. The difference between the two figures (FFED) provides an estimate of excess folding free energy across the length of the virus genome which was shown to be indicative of sequence dependent RNA structure.

The evolutionary conservation of potential sequence dependent RNA structures was assessed through folding representative examples of each of the major genotypes for both HCV (genotypes 1-6) and HGV/GBV-C (genotypes 1-4). However, the diversity between HGV/GBV-C genotypes is much less than that observed for HCV

(approximately 14% compared to 31-34%). Consequently GBV-A and GBV-B were analysed as out-groups, as they have similar genome organisations to HG/GBV-C and HCV respectively, yet show much greater divergence (approximately 40%). A recently isolated chimpanzee homologue (HGV/GBV-C_{CPZ}), which shows intermediate divergence between HGV/GBV-C and GBV-A was also analysed as a further out-group.

Prior to analysis the complete genome sequences of HCV (genotypes 1-6), HGV/GBV-C (genotypes 1-4), GBV-A and GBV-B were aligned by hand using the Simmonic 2000 package (Simmonds and Smith, 1999b), (See appendix for a list of sequences examined). The virus genomes were then split into 498 base fragments overlapping by 249 bases, in order to retain codon integrity, and 50 independent nucleotide randomisations made (randomisation methods explained in section 3.2.1). The FFE of each randomised sequence was then compared to that of each native sequence in order to make an estimate of FFED.

All FFE calculations were determined using MFOLD 3.1 (http:www.bioinfo.rpi.edu/~zukerm/) using the implementation available in the program ZIPFOLD (Zuker, 2003). MFOLD calculates the minimum FFE of a given sequence based on the sum of the contribution of nucleotide pairings, stacking and stem loop lengths (Rivas and Eddy, 2000; Zuker, 2000). The contribution of each nucleotide pairing (the "nearest neighbour" rules) to the overall free energy of a sequence is derived from experimental melting data (Mathews et al., 1999). MFOLD does not take into account higher order structures such as pseudoknots.

3.2 RESULTS

In order to investigate the existence of RNA structure within the polyprotein coding regions of HCV (genotypes 1-6) and HGV/GBV-C (genotypes 1-4) a comparison of FFE with those of sequence order randomised controls was made (Tuplin et al., 2002; Cuceanu et al., 2001). As a comparison GBV-B and GBV-A were also analysed, as they are divergent from HCV and HGV/GBV-C yet maintain similar structural organisation (Leary et al., 1996; Linnen et al., 1996).

3.2.1 SEQUENCE ORDER RANDOMISATION

It has previously been proposed that local base composition heterogeneity and biased dinucleotide distribution may result in artefactualy high FFEDs that are not the result of a sequence that evolved to produce stable RNA structure (Workman and Krogh, 1999; Rivas and Eddy, 2000). Consequently, in this study we developed and used a number of the sequence order randomisation methods in order to overcome these compounding effects (Tuplin et al., 2002).

1. **Nucleotide order randomisation (NOR)**: Nucleotide order was randomised with no account of local nucleotide or codon composition. This is the standard method used in most previous studies.
2. **Codon order randomisation (COR)**: Codon order was randomised, thus avoiding the disruption of sequence order within triplets.

3. **Like codon randomisation (CLR)**: Randomisation of codons for each specific amino acid. Following randomisation the amino acid sequence remained the same.

4. **Like codon swap (CLS)**: Pairwise exchange of neighbouring codons specifying each specific amino acid; may only be applied twice to a given native sequence. Following randomisation the amino acid sequence remained the same and there was little disruption of local nucleotide biases.

5. **Dinucleotide randomisation (CDR)**: Randomisation of each set of codons with identical first and third bases. Following randomisation the dinucleotide sequence remained the same but the amino acid sequence was changed.

6. **Dinucleotide swap (CDS)**: Pairwise exchange of neighbouring codons with identical first and third bases; may only be applied twice to a given native sequence. Following randomisation the dinucleotide sequence remained the same, the amino acid sequence was changed and there was little disruption of local nucleotide biases.

The difference in FFE of a native sequence to a set of randomised sequences has been shown to approximate to a normal distribution, with mean of zero and a standard deviation of + and -1 (Workman and Krogh, 1999; Rivas and Eddy, 2000). Consequently, the FFED between the native and randomised sequences was expressed as a Z-score (Workman and Krogh, 1999), which is the number of standard deviations between the free energy on folding the native sequence and the mean of the randomised sequences. This method took into account the range of

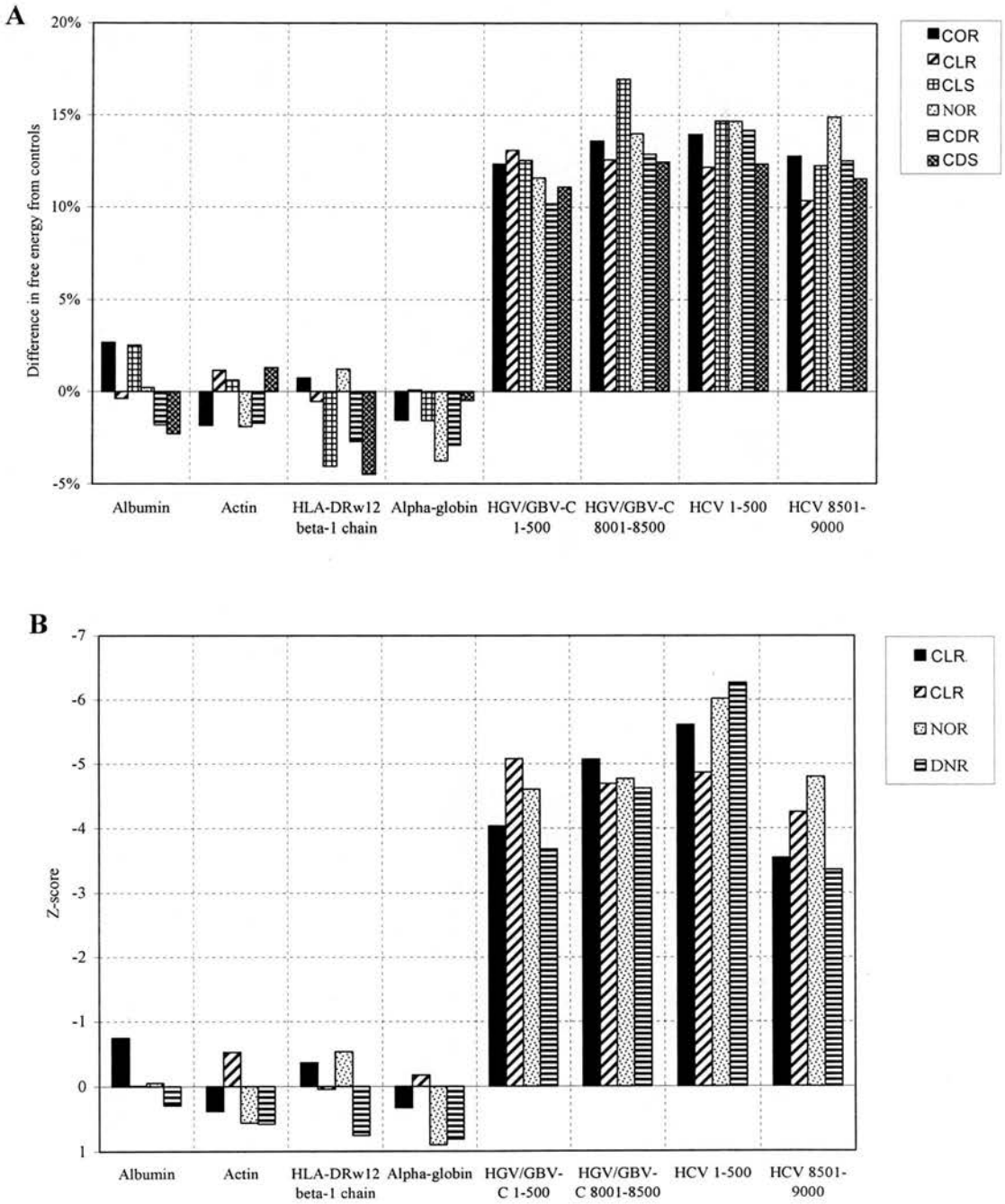
values between independent randomisations as well as just the difference in folding free energy to the native sequence. Z-scores of less than -2.33 were taken as being statistically significant at 1% for sequence order dependent RNA structure (Workman and Krogh, 1999). It was possible to determine Z-scores using the randomisation methods NOR, COS, CLR and CDR, as each was able to produce a multiple set of independent randomised sequences (Fig. 3.1). The nucleotide swapping methods CLS and CDS only generate two randomised sequences, preventing the calculation of a Z-score (Fig. 3.1).

3.2.2 CONTROL FOLDING FREE ENERGY INVESTIGATIONS

In order to investigate the effect of the six sequence order randomisation methods on FFED each method was applied to a range of mammalian sequences with no known or likely RNA secondary structure (the coding regions of mammalian albumin, actin, HLA class II and alpha-globin genes) (Fig. 3.1) (Cuceanu et al., 2001; Tuplin et al., 2002). These were compared to sequences in which RNA secondary structure had previously been demonstrated by analysis of covariant substitutions and reduction in synonymous variability (HGV/GBV-C coding regions 1-500 and 8001-8500 (Simmonds and Smith, 1999b; Cuceanu et al., 2001); HCV coding regions 1-500 and 8501-9000 (Han and Houghton, 1992; Smith and Simmonds, 1997a; Hofacker et al., 1998; Tuplin et al., 2002)) (Fig. 3.1).

Each of the six randomisation methods produced comparable FFEDs between the native and randomised sequences. None of the results for the mammalian sequences

Figure 3.1. **A:** Differences in folding free energy between native mammalian (albumin, actin, HLA class II, alphaglobin) and viral coding sequences (5' and 3' ends of HGV/GBV-C and HCV) and those between six different randomisation methods (NOR, COR, CLR, CLS, CDR and CDS). **B**: Corresponding Z-scores for NOR, COR, CLR and CDR.



produced a positive difference above 3% (Fig. 3.1A). For those methods from which Z-scores could be calculated (NOR, COR, CLR and CDR), the values were limited to above -1, indicating no significant FFEDS between the randomised and native sequences (Fig. 3.1B). Large differences were observed between the randomised and native sequences for both regions of HGV/GBV-C and HCV (Fig. 3.1). For both the 5' and 3' coding regions of HGV/GBV-C analysed the FFED by each of the six methods ranged from 10 to 17% (Fig. 3.1A); Z-scores ranged from -3.7 to -5.1, indicating a significant difference in folding free energy between the randomised and native sequences (Fig. 3.1B). Similar results were obtained for the 5' and 3' coding regions of HCV, with difference for each of the six methods ranging from 10 to 15% (Fig. 3.1B); Z-scores ranged from -3.4 to -6.3, again indicating a significant FFED between the randomised and native sequences (Fig. 3.1B).

These results provide evidence for sequence order dependent RNA secondary structure within the 5' and 3' coding regions of both HGV/GBV-C and HCV (Tuplin et al., 2002). They are also comparable to those observed for other viruses with well defined secondary structure such as plant viroids and the non-coding region of hepatitis delta virus; in which free energy differences of between 15% and 20% were observed, using the NOR method of sequence randomisation (Cuceanu et al., 2001). The similarity in excess folding free energy between the randomised and native sequence (and Z-scores where calculable) using the six different methods indicated that disruption of codon composition, dinucleotide frequencies or local differences in base composition did not account for the differences observed in the viral sequences (Tuplin et al., 2002).

3.2.3 FOLDING FREE ENERGY DIFFERENCES ALONG THE COMPLETE HCV POLYPROTEIN CODING REGION

As each of the six methods of calculating excess FFEDs gave comparable results the two least disruptive methods, that allowed Z-scores to be calculated, were used (CLR and CDR) for more detailed analysis of complete virus genomes. The polyprotein coding regions of HCV, genotypes 1a, 2a, 3a, 4a, 5a and 6a, were split into 498 base fragments overlapping by 249 bases (36 fragments over an alignment length of 9168 bases). Excess free energy on folding was compared between each native sequence and a set of 50 independently randomised sequence replicates, using both the CLR and CDR methods (Fig. 3.2). In each of the genotypes FFEDs between 6.2% and 8.8% was observed across the polyprotein coding region between the native and randomised sequences (mean values CLR: 7.8%, CDR: 6.9%; mean Z-scores: CLR: -2.6, CDR: -2.3).

In order to localize potential sequence dependent RNA secondary structure to defined regions of the HCV genome the mean values of each of the six genotypes were plotted against genome position (Fig. 3.3A). The greatest FFEDs were observed towards the 5' (fragments 1-498 and 251-749) (CLR; 9.5%, CDR: 10.5%; mean Z-scores: CLR -4.1, CDR -3.5) and 3' (fragments 7251-7749 and onwards) (CLR; 11.8%, CDR: 10.7%; mean Z-scores: CLR -3.5, CDR -3.9) extremes of the polyprotein coding region.

In order to investigate possible sequence dependent secondary structure within the replication intermediate of the HCV genome parallel testing of the reverse

Figure 3.2: Mean differences in folding free energy of 498-base fragments spanning viral genome of HCV genotypes 1a-6a, and GBV-B, using two randomisation methods (column 1 CLR and column 2 CDR). For each sequence, columns 1 and 3 correspond to CLR; 2 and 4 to CDR. For each sequence, columns 1 and 2 correspond to native sequence; 3 and 4 to reverse compliment. Z-scores indicated by shading.

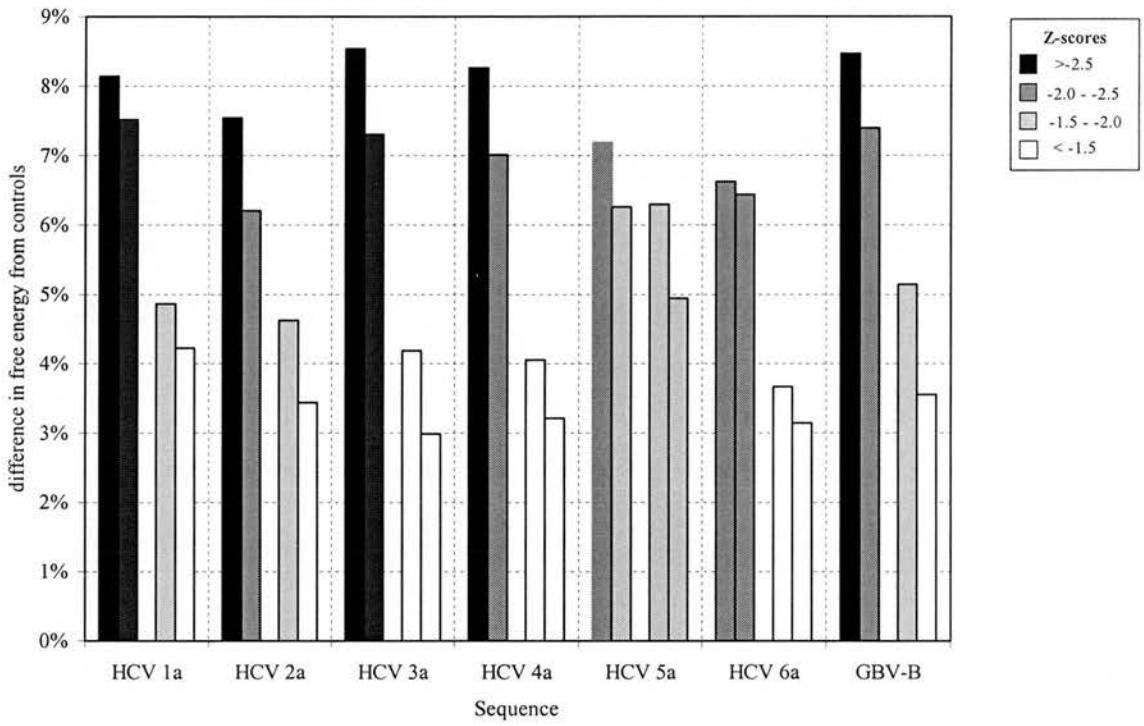


Figure 3.3: **A**: Mean difference in folding free energy of 498-base overlapping fragments of HCV genotypes 1-6, across different regions of the HCV genome using two randomisation methods (column 1 CLR; column 2 CDR). Z-score ranges are indicated by shading.

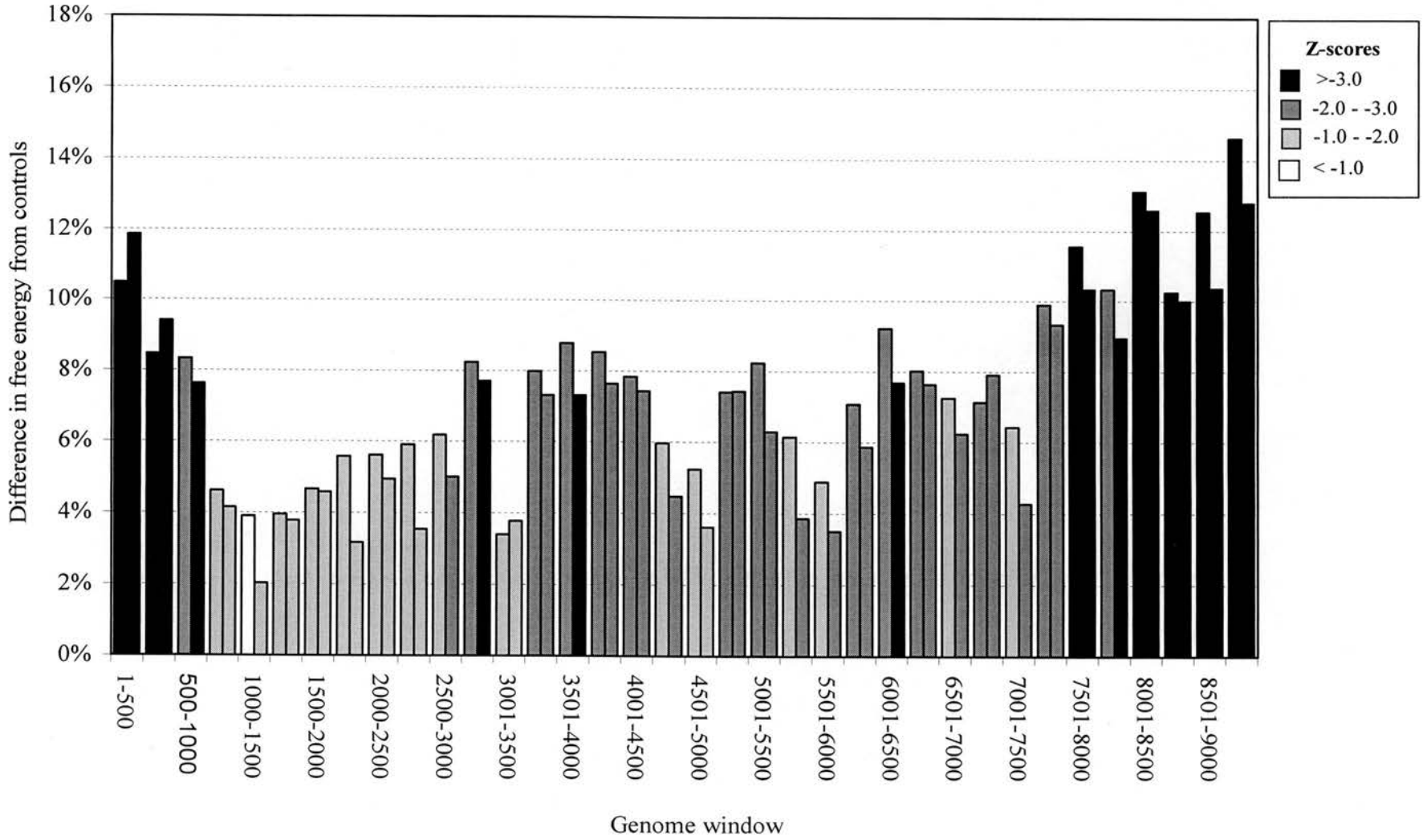
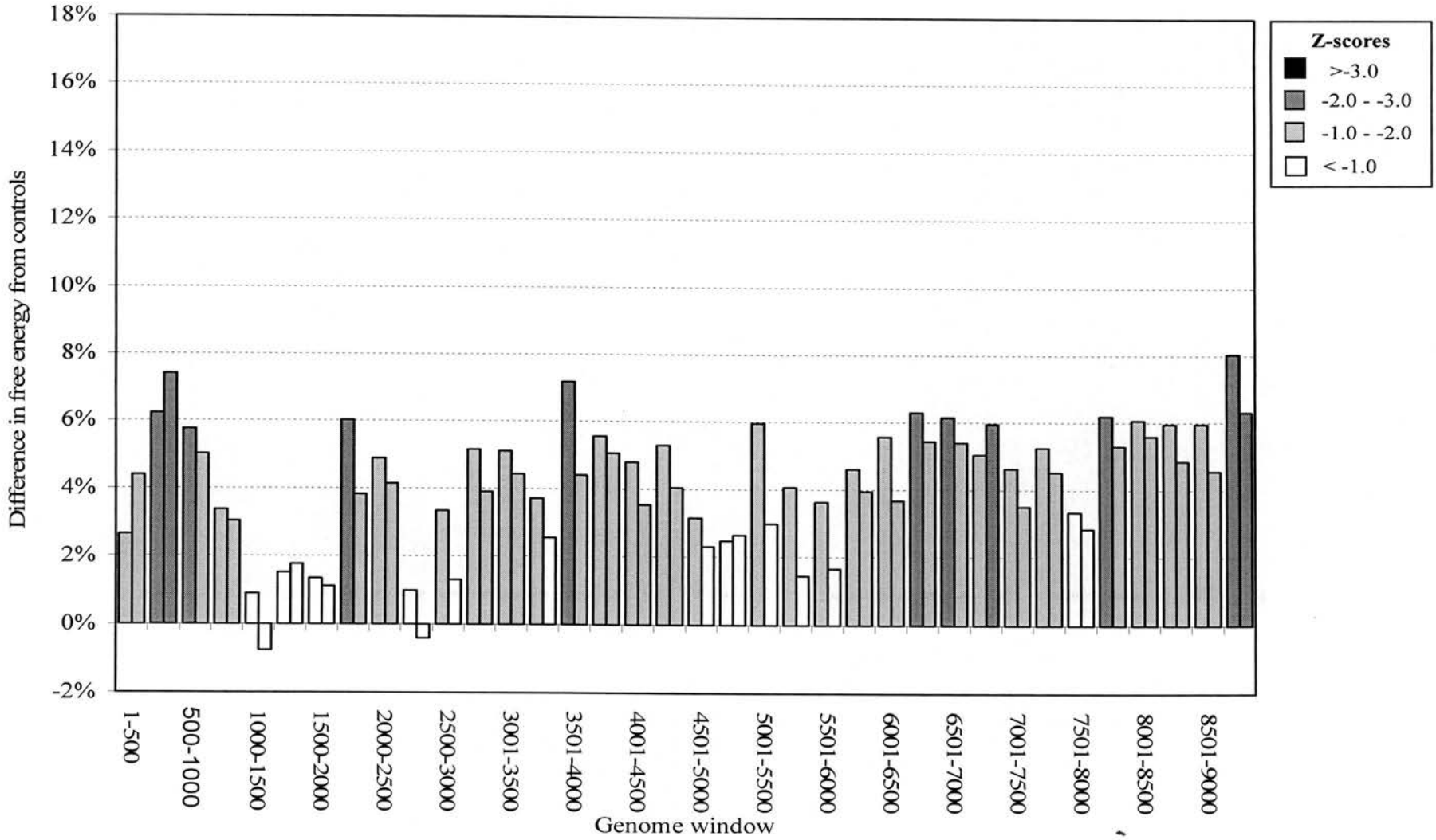


Figure 3.3. B: Free energy differences and Z-scores of corresponding reverse complement sequences across the HCV genome (column 1 CLR; column 2 CDR).



compliment sequence was performed. The antisense genomes showed consistently lower FFEDs between randomised and native sequences than was observed for their sense orientation equivalents (Figs. 3.3B). In each of the six genotypes examined levels of FFEDs between 3.0% and 6.3% were observed across the antisense polyprotein coding region (mean values CLR: 4.5%, CDR: 3.7%; mean Z-scores: CLR:-1.5, CDR:-1.2). This low excess free energy on folding was observed across all windows of the antisense genome. Values only consistently rose above 6%, in two fragments towards the 5' (fragment 251-749) (CLR; 6.2%, CDR: 7.4%; mean Z-scores: CLR -2.1, CDR -2.5) and 3' (fragment 8670-9168) (CLR; 8.0%, CDR: 6.3%; mean Z-scores: CLR -2.7, CDR -2.0) ends of the antisense polyprotein coding region. In no fragments were Z-scores for both methods within the level of statistical significance for the HCV replication intermediate.

3.2.4 FOLDING FREE ENERGY DIFFERENCES ALONG THE COMPLETE GBV-B POLYPROTEIN CODING REGION

GBV-B exhibited a similar levels of FFEDs between the native and randomised sequences as that observed for HCV (Fig. 3.2 and 3.4A), with a mean excess across the polyprotein coding region of 8.5% using the CLR method and 7.4% using CDR (Z-scores: CLR -2.6, CDR -2.3). The greatest differences were again observed towards the 5' (fragments 1-498 to 751-1249) (CLR; 11.1%, CDR: 9.7%; mean Z-scores: CLR -3.8, CDR -2.9) and 3' (fragments 7001-7499 and 7751-8249) (CLR;

Figure 3.4A: Mean difference in folding free energy of 498-base overlapping fragments of GBV-B across different regions of the virus genome using two randomisation methods (column 1 CLR; column 2 CRD). Z-score ranges are indicated by shading.

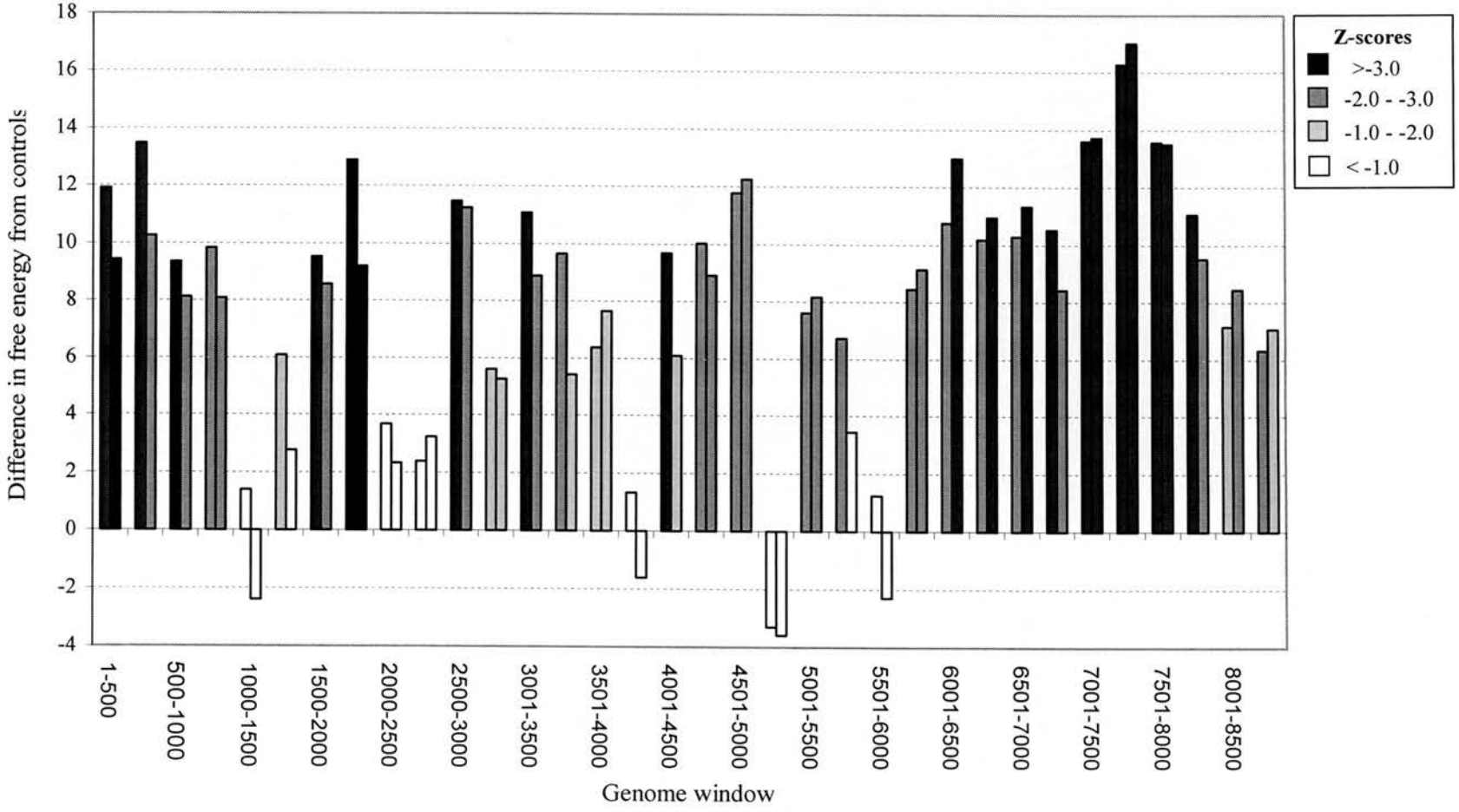
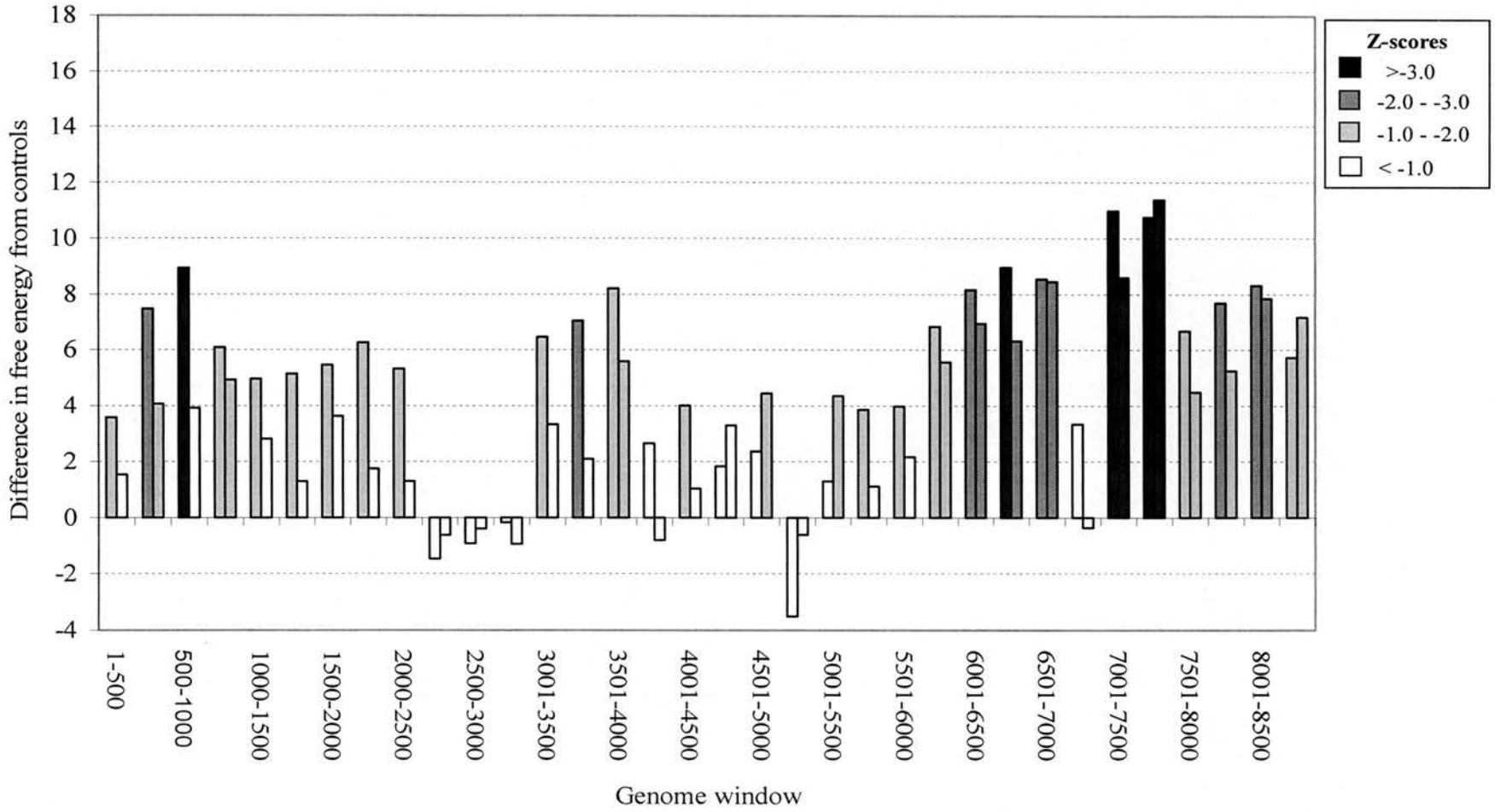


Figure 3.4B: Free energy differences and Z-scores of corresponding reverse, complement sequences across the GBV-B genome (column 1 CLR; column 2 CRD).



13.7%, CDR: 13.4%; mean Z-scores: CLR -2.2, CDR -3.9) extremes of the polyprotein coding region.

The reverse compliment replication intermediate of GBV-B showed much lower levels of FFED than the positive sense sequence (Fig. 3.4B), with a mean excess of 5.1% using the CLR methods and 3.6% using CDR (Z-scores: CLR -1.5, CDR -1.1). Low FFEDs were observed across all windows of the antisense genome, with values only rising above 6%, for both methods, towards the 3' extreme of the sequence. Two fragments within this region (7000-7500 and 7250-7750) consistently exhibited both FFEDs above 8% and statistically significant Z-scores (mean Z-scores: CLR -3.4, CDR -3.2).

3.2.5 FOLDING FREE ENERGY DIFFERENCES ALONG THE COMPLETE HGV/GBV-C AND GBV-A POLYPROTEIN CODING REGION

The polyprotein coding regions of HGV/GBV-C, genotypes 1, 2, 3 and 4, and GBV-A were divided into 35 and 36 fragments respectively, each measuring 498 bases in length and overlapping by 249 bases. The FFEDs of each sequence were analysed using the CLR and CDR algorithms, following the same methodology as previously used for HCV and GBV-B.

In each of the HGV/GBV-C genotypes FFEDs of between 9.5% and 13.2% were observed across the polyprotein coding regions (mean values CLR: 11%, CDR: 12.1%; mean Z-scores: CLR: -3.7, CDR: -4.1) (Fig. 3.5). Differences were consistently high across the length of the genome (Fig. 3.6A). However, the greatest

Figure 3.5. Mean differences in folding free energy of 498-base fragments spanning viral genome of HGV/GBV-C genotypes 1-4, and four examples of GBV-A, using two randomisation methods (CLR and CDR). For each sequence, columns 1 and 3 correspond to CLR; 2 and 4 to CDR. For each sequence, columns 1 and 2 correspond to native sequence; 3 and 4 to reverse complement. Z-scores are indicated by shading.

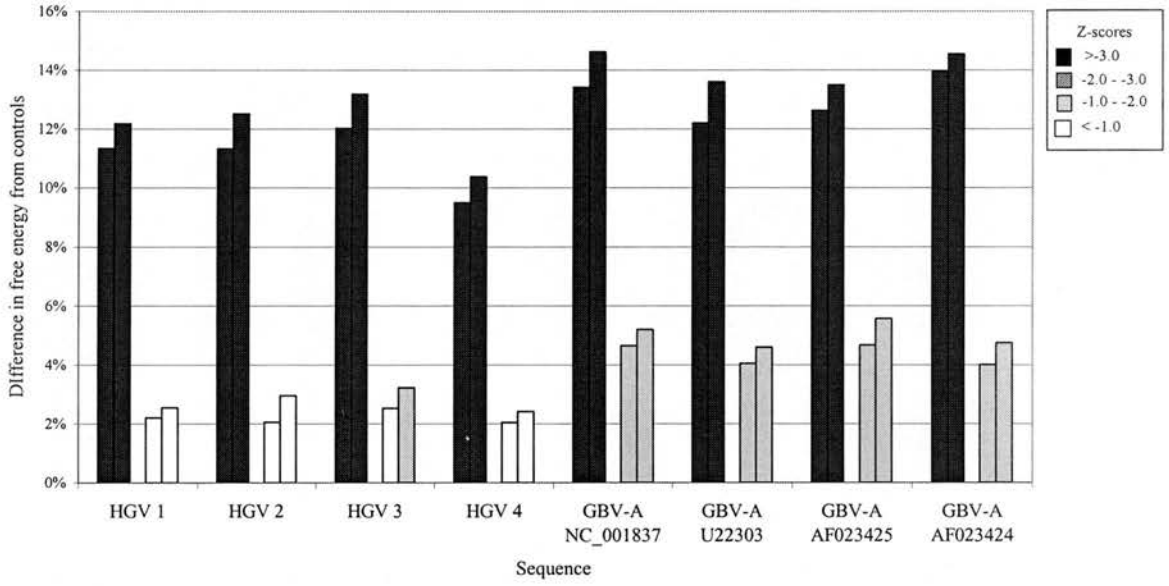


Figure 3.6. **A**: Mean difference in folding free energy of 498-base overlapping fragments of HGV/GBV-C genotypes 1-4 across different regions of the virus genome using two randomisation methods (column 1 CLR; column 2). Z-score ranges are indicated by shading.

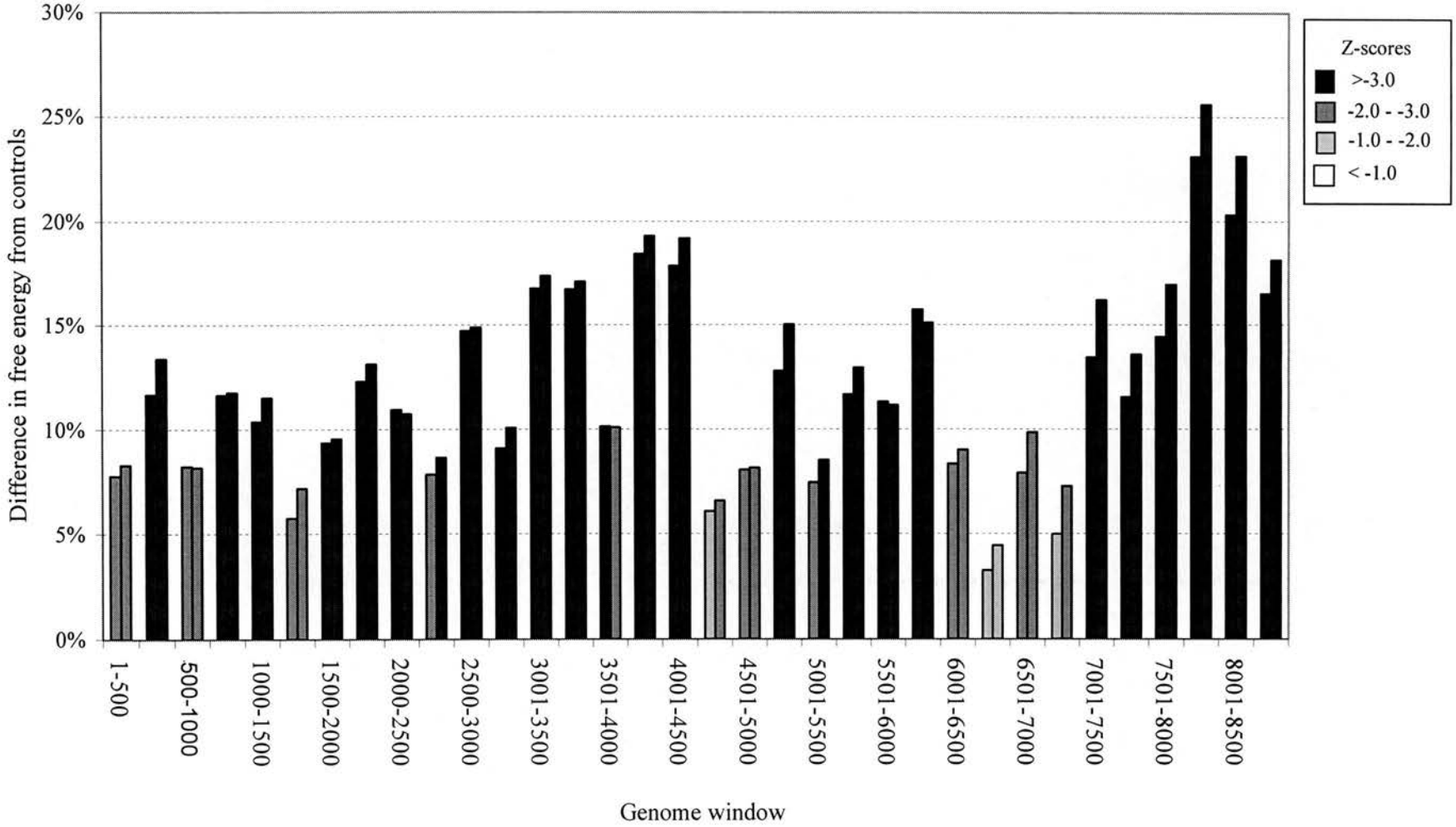
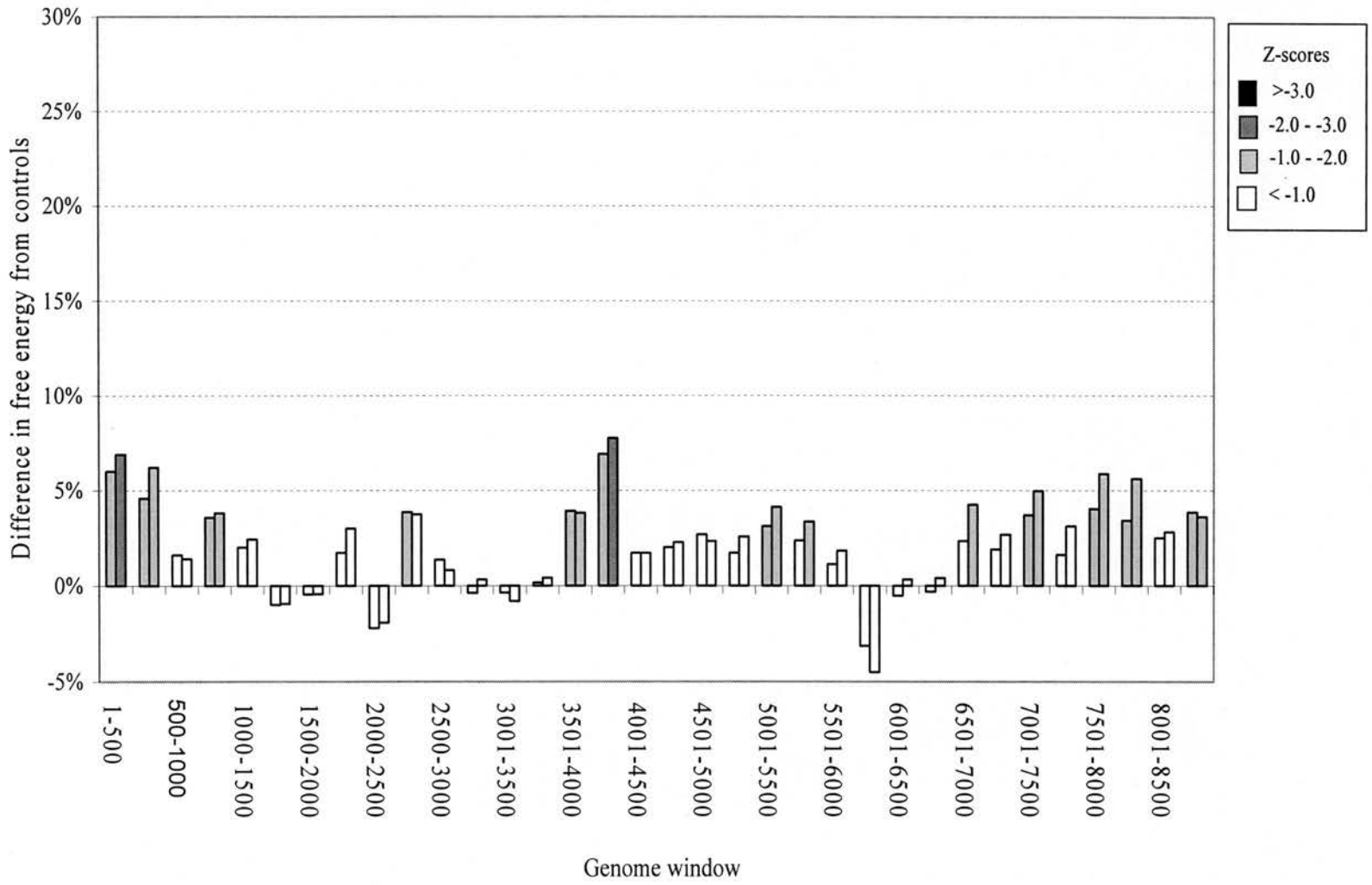


Figure 3.6. B: Free energy differences and Z-scores of corresponding reverse, complement sequences across the HGV/GBV-C genome. (column 1 CLR; column 2 CDR)



differences were observed towards the midpoint (fragments 3001-3499 to 4001-4499) (CLR; 16%, CDR: 16.6%; mean Z-scores: CLR -4.9, CDR -5.2) and 3' (fragments 7751-8249 to 8251-8749) (CLR; 19.9%, CDR: 22.3%; mean Z-scores: CLR -5.5, CDR -6.4) regions of the polyprotein coding region.

Contrasting results were obtained for the HGV/GBV-C replication intermediate sequences, in which low levels of FFED were again observed in all genotypes examined, ranging from 2.1% to 3.2 % (Fig. 3.5). Differences were consistently low along the length of the reverse compliment sequence (Fig. 3.6B). In only three fragments were FFEDs above 5% (fragments 1-498, 3751-4249 and 8498-8754). Fragment 8256-8754 was the only one in which a significant Z-score was observed (CLR; 7.8%, CDR: 8.2%; mean Z-scores: CLR -2.3, CDR -2.4).

Similar results were observed for GBV-A sequences examined with FFEDs between 12.2% and 14.6% noted across the polyprotein coding region (mean values CLR: 13.1%, CDR: 14.1%; mean Z-scores: CLR: -4.2, CDR: -4.4) (Fig. 3.5). As with HGV/GBV-C, levels of FFED were consistently high across the length of the genome (Fig. 3.7A), with the greatest differences observed towards the midpoint (fragment 4001-4498) (CLR; 18.8%, CDR: 19.8%; mean Z-scores: CLR -5.7, CDR -5.8) and 3' (fragments 8001-8598 to 8614-9112) (mean values CLR: 20.9%, CDR: 22.4%; mean Z-scores: CLR: -6.2, CDR: -6.4) regions of the GBV-A polyprotein coding region.

Again contrasting results were obtained for GBV-A replication intermediate sequences, with low levels of FFEDs, ranging from 4.0% to 5.6%, along the length of antisense sequence (Fig. 3.5 and 3.7B). However, towards the 3' extreme of the sequence fragments with differences above 5% and significant Z-scores were

Figure 3.7. **A**: Mean difference in folding free energy of 498-base overlapping fragments of GBV-A across different regions of the virus genome using two randomisation methods (column 1 CLR; column 2 CDR). Z-score ranges are indicated by shading.

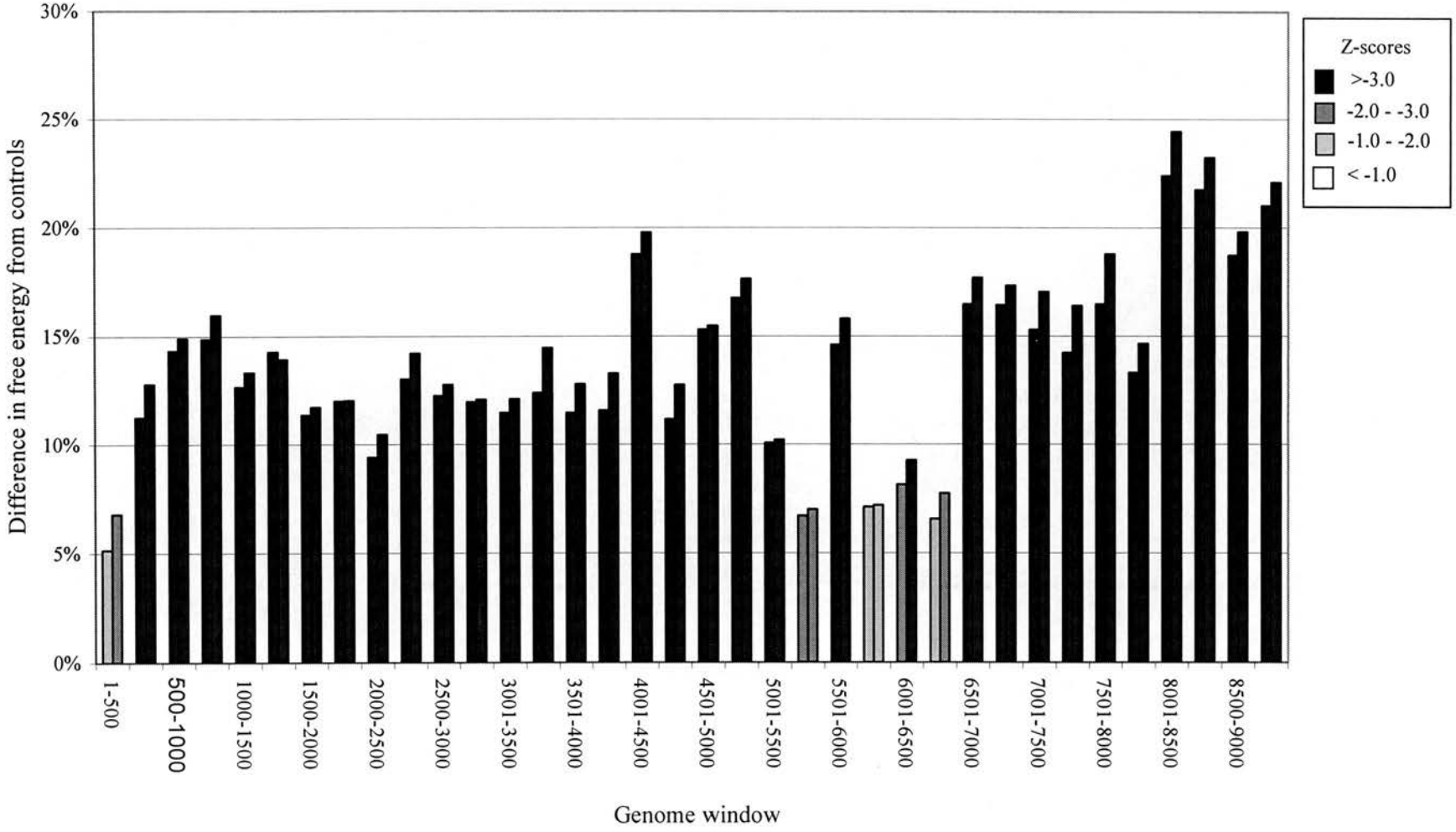
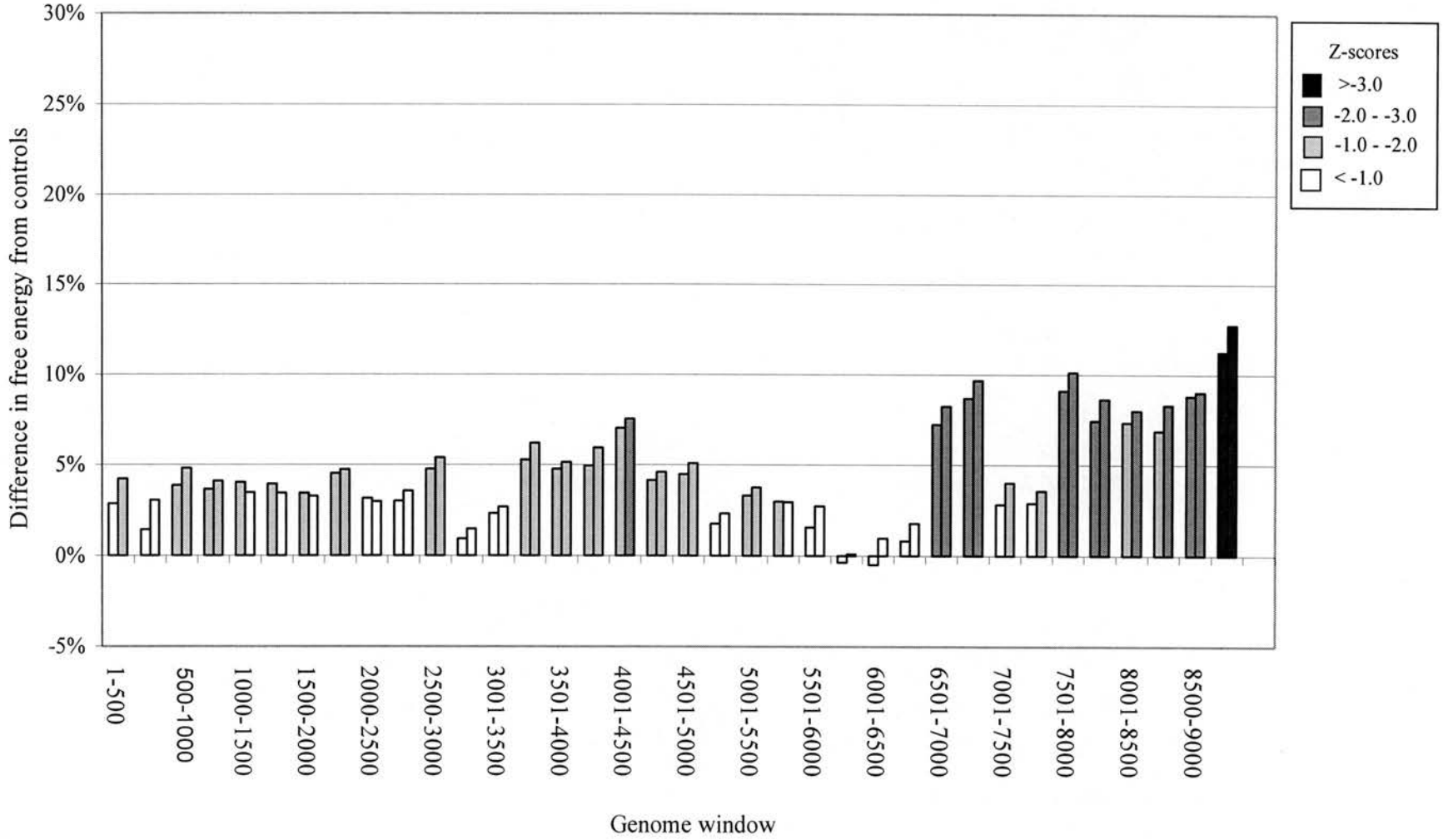


Figure 3.7. B: Free energy differences and Z-scores of corresponding reverse, complement sequences across the GBV-A genome (column 1 CLR; column 2 CDR).



observed (fragments 6501-6999, 6751-7249, 7501-7999, 7751-8249, 8501-8999 and 8751-9112). Fragment 8614-9112 was of note as it had an FFED above 10% and a Z-score below -3.0 (CLR: 11.2%, CDR: 12.7%; Z-scores: CLR: -3.2, CDR: -3.1).

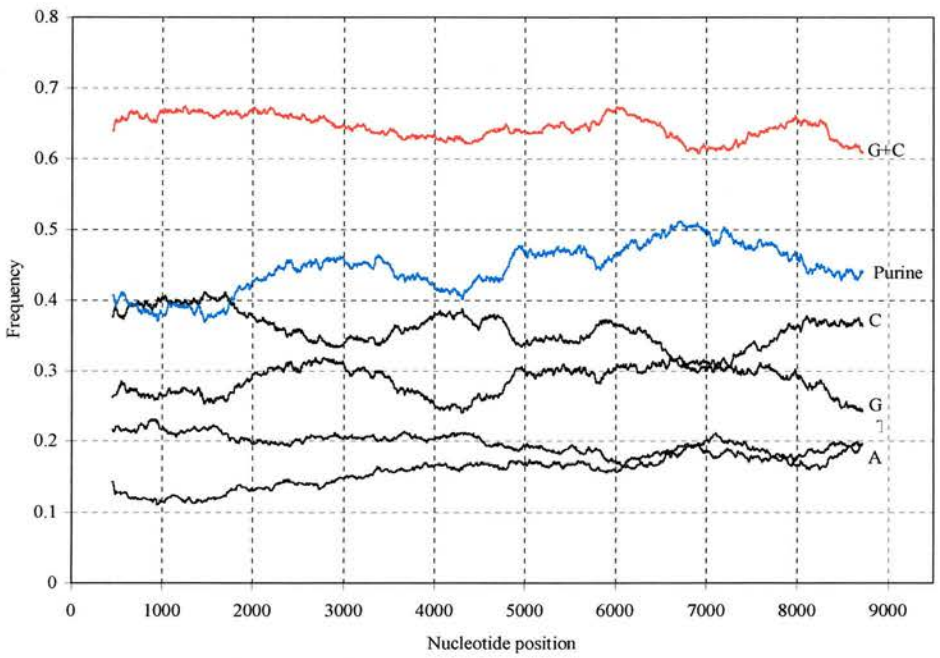
In summary, these results suggest that sequence dependent RNA secondary structure is distributed throughout the genomes of HCV and HGV/GBV-C as well as the related viruses GBV-B and GBV-A. The contrasting results observed between the positive and negative sequence orientations suggest that RNA structure is most relevant biologically for the virus RNA in its positive sense orientation.

3.3 DISCUSSION

In this study we have used excess free energy on folding in order to predict the existence of extensive sequence dependent RNA secondary structure across the genomes of both HCV and HGV/GBV-C and the closely related viruses GBV-B and GBV-A (Cuceanu et al., 2001; Tuplin et al., 2002). The results concur closely with those obtained using phylogenetic methods in which we observed a reduction in synonymous variability and clustering of covariant substitutions within the polyprotein coding regions of both viruses (Smith and Simmonds, 1997a; Simmonds and Smith, 1999b; Cuceanu et al., 2001; Tuplin et al., 2002).

Previously published analysis of prokaryotic and eukaryotic secondary structure using excess FFE has revealed a number of potential problems associated with the randomisation of local nucleotide composition which may lead to artefactual results. These include the effect of GC rich regions (GC islands) which may have a relatively high free energy on folding when compared with the randomised controls where any local nucleotide biases have been homogenised throughout the sequence (Rivas and Eddy, 2000). We were able to discount local nucleotide heterogeneity as a contributing factor in HCV as the base composition was shown to be relatively homogenous throughout the genome with an over representation of G/C residues at the third base position (Fig. 3.8) (Tuplin et al., 2002). We also developed a codon swapping method (CLS) of randomisation that minimised the distance between shuffled sites, thus maintaining local nucleotide biases. CLS gave comparable results to the other randomisation methods, for virus sequences examined, which further discounted sequence heterogeneity as a compounding factor.

Figure 3.8: Scan of base composition at third codon positions across the polyprotein coding region of HCV genotypes 1-6 (mean values), including combined values for (G + A) and (G + C).



A second cause of artefactualy high differences in folding free energy results from the disruption of dinucleotide frequencies (Workman and Krogh, 1999). The main contribution to free energy on folding an RNA molecule is associated with base pairing. However, a second contributing factor is the two neighbouring residues to the paired bases. For example, a C-G base pair is more thermodynamically favourable than a G-C base pair when stacked on top of an A-U base pair (Workman and Krogh, 1999). Consequently, a randomisation algorithm which disrupts dinucleotide frequencies may lead to large differences in FFE, between the native and randomised sequences, which are independent of RNA secondary structure.

Excess FFE, utilising a dinucleotide disrupting algorithm, was used in a previous study to predict stable RNA structure in 51 prokaryotic and eukaryotic mRNAs (Seffens and Digby, 1999). It has since been shown by Workman and Krogh that in the majority of the mRNAs examined the excess free energy on folding was an artefact, due to the disruption of dinucleotide frequencies (Workman and Krogh, 1999). In this study we developed two randomisation methods that retained the dinucleotide frequencies and codon structure of the native sequences (CDR and CDS); CDS also retains any local differences in dinucleotide frequencies. Further evidence that unusual dinucleotide frequencies were not responsible for the folding free energy differences observed in HCV or HGV/GBV-C was provided by the observation of only very limited dinucleotide frequency biases in either virus. Of sixteen dinucleotide pairs only the frequencies of CG and UG differed from those expected (Tuplin et al., 2002). However, these differences were not localised to either the 5' or 3' extremes of either virus genome, in which the greatest differences in folding free energy were observed.

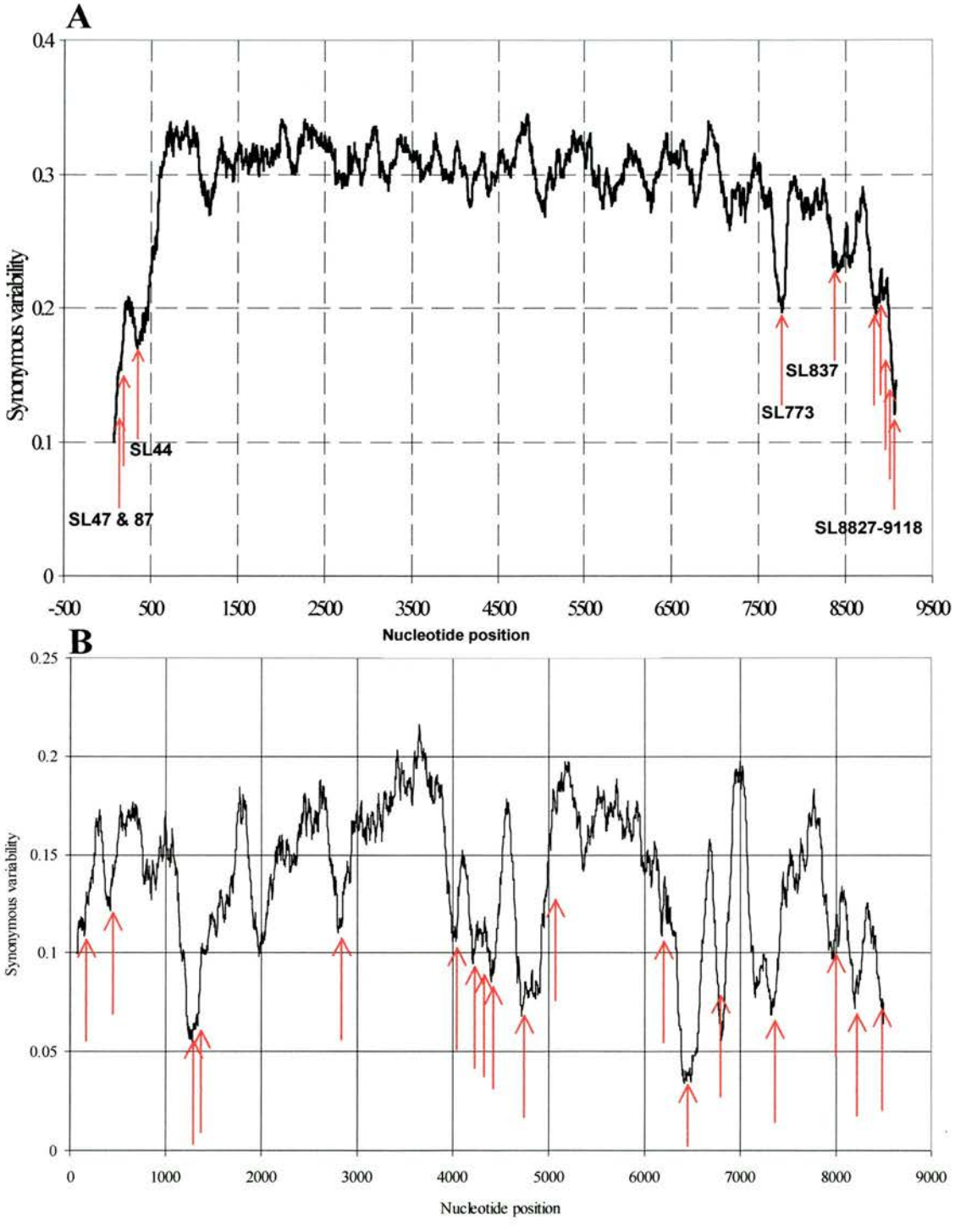
In order to determine the relationship between the six different randomisation algorithms developed and FFEDs, each of the methods was applied to the coding regions of a number of mammalian genes, with no known or likely RNA secondary structure (Fig. 3.1). They were also used to examine the extreme upstream and downstream fragments of HCV and HGV/GBV-C, in which RNA secondary structure had previously been predicted by independent methods (Fig. 3.1). Each of the randomisation methods gave comparable results indicating the presence of sequence dependent RNA structure within the extreme 5' and 3' polyprotein coding region of both viral genomes and a lack of structure in each of the mammalian sequences examined. The folding free energies observed within the virus sequences were comparable to those observed for plant viroids and the non-coding region of delta virus (Cuceanu et al., 2001), in which RNA structure is known to play a functional role during genome replication. The results were also comparable to those obtained in a study by Workman and Krogh, where they observed that large, functional, highly structured RNA molecules, such as rRNA, possess a high excess in FFE which is comparable to that observed for the viruses examined in this study (mean Z-score -7.93, using a randomisation method which maintained dinucleotide frequencies) (Workman and Krogh, 1999). Further evidence for the veracity of the results observed in this study is provided by the fact that no excess folding free energies were observed within the mammalian mRNA sequences even though they possesses a wide diversity of sequence compositions and dinucleotide biases.

Analysis of the complete coding regions of HCV, HGV/GBV-C, GBV-B and GBV-A revealed excess FFE across the length of the genomes; with the greatest excesses towards the 5', middle and 3' domains of the coding regions of HCV,

HGV/GBV-C and GBV-A. In GBV-B the largest excesses were also observed towards the 5' and 3' extremes of the coding region, yet towards the middle of the virus genome the differences were more sporadic.

We have previously analysed sequence alignments of all available epidemiologically unlinked HCV (Smith and Simmonds, 1997a; Tuplin et al., 2002), and HGV/GBV-C (Simmonds and Smith, 1999b) (Cuceanu et al., 2001), sequences for the presence of covariant substitutions and the frequency of synonymous substitutions. Reductions in sequence diversity at synonymous sites and the existence of covariant substitutions may result from constraints on sequence change imposed by RNA secondary structure. A very close concordance was observed between fragments which possessed large FFEDs and regions in which such constraints in sequence divergence were observed for both HCV and HGV/GBV-C (Fig. 3.9). In HCV the greatest reductions in synonymous variability and excesses in FFE were observed in the most extreme 5' and 3' fragments, in which the greatest clustering of covariant substitutions were also observed. As with excess FFE, suppression of synonymous variability and covariant substitutions were more extreme and dispersed further across the genome of HGV/GBV-C than was the case for HCV. However, a close agreement in position was again noted, especially towards the 3' of the genome within the NS5 coding regions (approximately down stream of position 7000) and towards the middle of the genome within the NS3/NS4 (4000-5000) coding regions. It has not been possible to complete such phylogenetic analysis for GBV-A or GBV-B due to the lack of comparative sequence data and genotypes. Agreement between these independent phylogenetic and thermodynamic methods adds further veracity to the FFE results presented in this study, which

Figure 3.9. **A**: Variability at synonymous sites across the genome of HCV, genotypes 1-6 (mean values shown); location of covariant sites indicated by red arrows and stem loop names labelled (chapter 4) (Smith and Simmonds 1997a; Tuplin et al 2002). **B**: Synonymous variability across the genome of HGV/GBV-C genotypes 1-4 (mean values shown); location of covariant sites indicated by a red arrow (Simmonds and Smith 1999b).



suggest that sequence dependent RNA structure is distributed throughout the genome of HCV, HGV/GBV-C, GBV-B and GBV-A.

In analysis of the reverse complement sequences large FFEDs were observed to be significantly reduced from those observed within the positive sense RNA genomes (Figs. 3.2 and 3.5). This suggests that RNA structure is more functionally relevant for the positive sense RNA genomes examined than the antisense replication intermediates. However, in both GBV-B and GBV-A discrete antisense fragments were observed with significant excess FFE (Figs. 3.4B and 3.6B). In both instances such fragments were observed within regions in which the greatest differences were also observed for the positive sense genomes. Due to non-canonical base pairing and different dinucleotide stacking energies the sense and antisense sequences would not be expected to possess the same folding free energies. However, it is likely that a sequence in which highly thermodynamically stable RNA structure were present would also posses a residual higher then expected free energy on folding in the opposing orientation, even though no biologically significant structure was present.

Using the same parameters we previously observed 14 covariant substitutions within HCV, compared to 48 within HGV/GBV-C, in which lower levels of suppression of synonymous variability were also noted (Simmonds and Smith, 1999b; Tuplin et al., 2002). It may be that the greater levels of sequence diversity between HCV genotypes prevented detection of less well conserved stem loops within HCV. However, the agreement between the phylogenetic and thermodynamic evidence is consistent with the existence of less RNA structure within the HCV genome than that of HGV/GBV-C. The absence of phylogenetic evidence for GBV-B and GBV-A makes it difficult to make a comparison with the other two viruses.

Nevertheless, based purely on the thermodynamic evidence, the level of RNA structure within GBV-B may be comparable to that observed in HCV and that of GBV-A to HGV/GBV-C which mirrors the genome organisation and phylogenetic relationship between the four viruses.

An even greater contrast is seen on analysis of other related viruses such as pestiviruses in which no thermodynamic evidence for RNA secondary structure within the polyprotein coding region was observed. Analysis of both BVDV and CSFV showed that the polyprotein coding regions of both viruses lack significant FFEDs when compared to randomised sequence controls (3.2% and 2.1% respectively) (Tuplin et al., 2002). Phylogenetic analysis has also failed to show any bias in synonymous variation or covariant substitutions (Peter Simmonds, personal communication).

In summary, we have used excess free energy to predict the existence of extensive sequence dependant RNA secondary structure across the polyprotein coding regions of a number of related single stranded RNA viruses (HCV, HGV/GBV-C, GBV-B and GBV-A). In doing so have developed a number of novel sequence randomisation methods in order to account for potential sequence composition heterogeneity.

CHAPTER 4

COMPUTATIONAL RNA STRUCTURE PREDICTION

4.1 INTRODUCTION

In chapter 3 large excesses in FFED were observed across the genomes of HCV and HGV/GBV-C, which were conserved between divergent genotypes and related viruses (GBV-B and GBV-A) and are consistent with sequence dependent RNA structure. Levels of FFED were comparable to those observed for viruses with well defined secondary structure such as plant viroids and Delta virus. Although, levels of FFED were high across the length of the genomes the greatest excesses were observed within core gene of HCV and the NS5B regions of HCV and HGV/GBV-C. A number of previous investigations have shown that these regions exhibit the greatest suppression of synonymous variability and clustering of covariant substitutions, which are consistent with restraints on sequence divergence due to functionally conserved RNA structure (more detail in chapter 1.9).

Based on the concordance of the previous phylogenetic and thermodynamic studies specific structural predictions are made in chapter 4 for the core gene and NS5B region of HCV and the NS5B region of HGV/GBV-C. Secondary structure predictions were also made for the 3'UTR of HGV/GBV-C as a comparison with related viruses, such as GBV-A and HCV, suggest that this region is likely to be highly structured.

Secondary structure predictions were made using MFOLD 3.1 with default setting (Zuker, 2003; Mathews et al., 1999). MFOLD predicts all possible secondary structure conformations into which an RNA molecule can fold, based on minimum

folding free energy; which is the sum of the contribution of nucleotide pairings, stacking energies and stem loop lengths (Rivas and Eddy, 2000; Zuker, 2000).

In order to assess the evolutionary conservation of predicted structures parallel folding of all epidemiologically unlinked, complete genome sequences available on GenBank was performed for both HCV (genotypes 1-6) and HGV/GBV-C (genotypes 1-4). As the sequence divergence between HGV/GBV-B is comparatively high (approximately 13%) the chimpanzee homologue HGV/GBV-C_{CPZ} was also analysed as an out-group, as it is has the same genome organisation as HGV/GBV-C but is relatively divergent (approximately 27%). Conservation was assessed between parallel foldings and included analysis of both covariant and semi-covariant substitutions.

SEQUENCES ANALYSED FOR RNA SECONDARY STRUCTURE

Complete genome sequences of both HCV and HGV/GBV-C (including HGV/GBV-C_{cpz}) were aligned by hand using the Simmonic 2000 package (Simmonds and Smith, 1999b), prior to analysis with MFOLD (see appendix for alignments and a list of GenBank accession numbers). Alignment of the HGV/GBV-C homologous nucleotides indicated that the terminal 55 nucleotides of the 3'UTR were missing from the chimpanzee isolate. The HCV sequence alignment was divided into the core gene including the extreme 3' of the 5'UTR (nucleotides -23 to 516) and the NS5B region (nucleotides 8717 to 9186). The last 1000 nucleotides of the HGV/GBV-C alignment were separated into the NS5B region and 3'UTR. The

NS5B region of HGV/GBV-C included the whole sequence upstream of and including the stop codon (nucleotides 8231 to 8901), whilst the 3'UTR included the whole sequence downstream of this.

4.2 RESULTS

Specific predictions of RNA secondary structure were made for regions of both HCV and HGV/GBV-C in which there was a suppression of synonymous variability, an excess free energy on folding as compared with sequence order randomised controls and a clustering of covariant substitutions (chapter 3). The regions investigated were the core gene and NS5B region of HCV (Tuplin et al., 2002) and the NS5B encoding region and 3'UTR of HGV/GBV-C (Cuceanu et al., 2001). Secondary structure predictions were made using the program MFOLD 3.1 with default settings (Mathews et al., 1999).

Conservation of each predicted RNA structure was assessed by parallel folding of all epidemiologically unlinked, complete genome sequences available on GenBank of HCV (genotypes 1-6) and HGV/GBV-C (genotypes 1-4 and HGV/GBV-Ccpz) and the retention of structural prediction between divergent genotypes, covariant and semi-covariant sites.

4.2.1 HCV SECONDARY STRUCTURE PREDICTION

Ten thermodynamically stable and evolutionarily conserved RNA structures were observed within the coding region of HCV; three within the core gene and seven within the NS5B region. The RNA structures were provisionally named according to the position within the HCV alignment of the first nucleotide at the 5' end of the base paired region. The core gene structures SL47, SL87 and SL443 spanned regions 47

to 84; 87 to 167 and 443 to 475 respectively (Fig. 4.1). The seven NS5B encoding structures SL7730, SL8376, SL8828, SL8926, SL9011, SL9061 and SL9118 spanned regions 7730 to 7777; 8376 to 8455; 8828 to 8897; 8926 to 8987; 9011 to 9058; 9061 to 9006 and 9118 to 9148 respectively (Fig. 4.2).

All the predicted RNA structures exhibited a high degree of conservation between genotypes and both covariant and semi-covariant substitutions (Table 4.1), (Fig. 4.1 and 4.2). However, some variation in structure length, sequence conservation in unpaired regions and predicted base pairings was observed. In predicted base-paired regions the third base positions were most commonly aligned with the downstream third codon position so that covariant sites were usually synonymous (in seven of the ten predicted stem-loops). Non-synonymous covariant substitutions were observed in SL47, SL8833, SL8926 and SL9118 resulting from base pairing between nucleotides at different codon positions.

Predicted HCV core gene RNA structures

The core gene stem-loops **SL47**, **SL87** and **SL443** were highly conserved between genotypes 1-6 (Fig. 4.1). However, there was variation in the length of SL47 within and between genotypes; in ~25% of sequences the first paired nucleotides were 48 and 83 as opposed to 47 and 84. A total of two covariant sites were observed within SL47, four within SL87 and five within SL443 (Table 4.1). There were a number of instances, in all three core gene stem loops, when a single substitution resulted in the destabilisation of individual paired nucleotides; in the majority of cases such

Figure 4.1. Schematic representation of HCV core gene conserved RNA structures. Covariant sites are highlighted in grey.

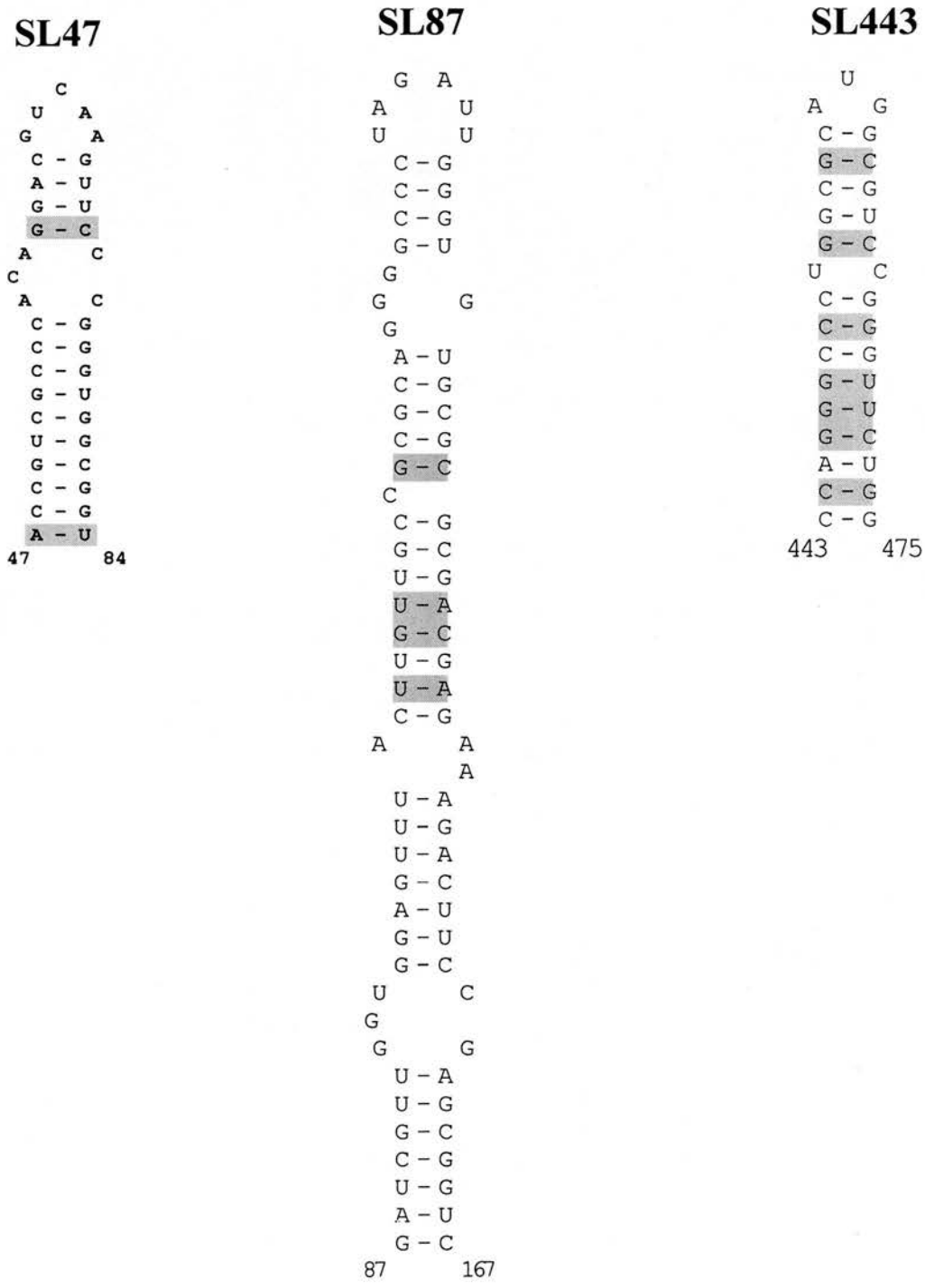


Figure 4.1: Schematic representation of HCV core gene and NS5B coding region conserved RNA stem loops. Stem loop names are given above the structures; positions of the upstream and downstream extent of base pairings are given below the structures and are numbered from the start of the HCV alignment (see appendix). Covariant sites are indicated by shading of paired nucleotides. Symbol '-' canonical Watson-Crick base pairing or G-U pairing. Examples of structurally different stem-loops (SL7730 and SL8376) indicated in boxes.

Table 4.1 Structure conservation of predicted stem loops within the core gene of HCV.

Stem-loop	genotype	N[u*]	+++ %	++ %	+ %	Co-variance	Stable variance †
SL47	1a	10	100	0	0	0	4/5
	1b	81	100	0	0	0	7/8
	2a	22	100	0	0	1	5/6
	3a	6	100	0	0	1	5/5
	4a	1	100	0	0	/	/
	5a	1	100	0	0	/	/
	6a	6	100	0	0	0	5/7
SL87	1a	10	100	0	0	1	8/10
	1b	81	100	0	0	0	8/9
	2a	22	100	0	0	0	7/9
	3a	6	100	0	0	2	7/12
	4a	1	100	0	0	/	/
	5a	1	100	0	0	/	/
	6a	6	100	0	0	4	7/9
SL443	1a	10	100	0	0	4	3/3
	1b	81	100	0	0	1	4/5
	2a	22	100	0	0	2	4/5
	3a	6	100	0	0	2	5/8
	4a	1	100	0	0	/	/
	5a	1	100	0	0	/	/
	6a	6	100	0	0	5	1/1

Structures were scored from + to +++ depending on the degree of similarity to the most common structure. +++ Stem-loop structurally identical; ++ minor differences in base pairing and/or size but the overall stem-loop conservation was maintained; + different structure in the same region. The frequency within which single synonymous substitutions retain or had a neutral effect on stem-loop structure is also shown. * Number of sequences examined for a given genotype: † Number of stabilising or neutral synonymous substitutions compared to the total number of synonymous substitution (semi-covariance).

Table 4.2: Structure conservation of predicted stem loops within the NS5B encoding region of HCV.

Stem-loop	genotype	N[u]*	+++ %	++ %	+ %	Co-variance	Stable variance†
SL7730	1a	10	100	0	0	1	5/7
	1b	81	100	0	0	1	6/6
	2a	22	100	0	0	0	2/3
	3a	6	100	0	0	2	5/5
	4a	1	100	0	0	/	/
	5a	1	100	0	0	/	/
	6a	6	100	0	0	2	3/4
SL8376	1a	10	100	0	0	4	7/11
	1b	81	100	0	0	0	11/19
	2a	22	100	0	0	2	10/15
	3a	6	0	100	0	3	7/11
	4a	1	0	100	0	/	/
	5a	1	0	100	0	/	/
	6a	6	0	100	0	3	/
SL8828	1a	10	90	10	0	0	13/14
	1b	81	100	0	0	0	15/20
	2a	22	100	0	0	1	8/12
	3a	6	100	0	0	1	11/12
	4a	1	100	0	0	/	/
	5a	1	100	0	0	/	/
	6a	6	/	16.7	83.3	/	/
SL8926	1a	10	60	40	0	2	12/17
	1b	81	100	0	0	1	8/15
	2a	22	100	0	0	1	6/6
	3a	6	/	100	0	0	8/12
	4a	1	/	100	0	/	/
	5a	1	/	100	0	/	/
	6a	6	/	100	0	/	/
SL9017	1a	10	100	0	0	2	5/8
	1b	81	100	0	0	1	2/2
	2a	22	100	0	0	3	9/9
	3a	6	100	0	0	5	4/4
	4a	1	100	0	0	/	/
	5a	1	100	0	0	/	/
	6a	6	100	0	0	5	6/7
SL9061	1a	10	100	0	0	2	7/7
	1b	81	100	0	0	1	4/4
	2a	22	100	0	0	1	2/2
	3a	6	100	0	0	3	7/7
	4a	1	100	0	0	/	/
	5a	1	100	0	0	/	/
	6a	6	100	0	0	2	3/3
SL9118	1a	10	100	0	0	1	4/4
	1b	81	100	0	0	1	5/6
	2a	22	100	0	0	1	9/10
	3a	6	100	0	0	0	3/3
	4a	1	100	0	0	/	/
	5a	1	100	0	0	/	/
	6a	6	100	0	0	0	3/3

Table 4.2: Structures were scored from + to +++ depending on the degree of similarity to the most common structure. +++ Stem-loop structurally identical; ++ minor differences in base pairing and/or size but the overall stem-loop conservation was maintained; + different structure in the same region. The frequency within which single synonymous substitutions retain or had a neutral effect on stem-loop structure is also shown. * Number of sequences examined for a given genotype: † Number of stabilising or neutral synonymous substitutions compared to the total number of synonymous substitution (semi-covariance).

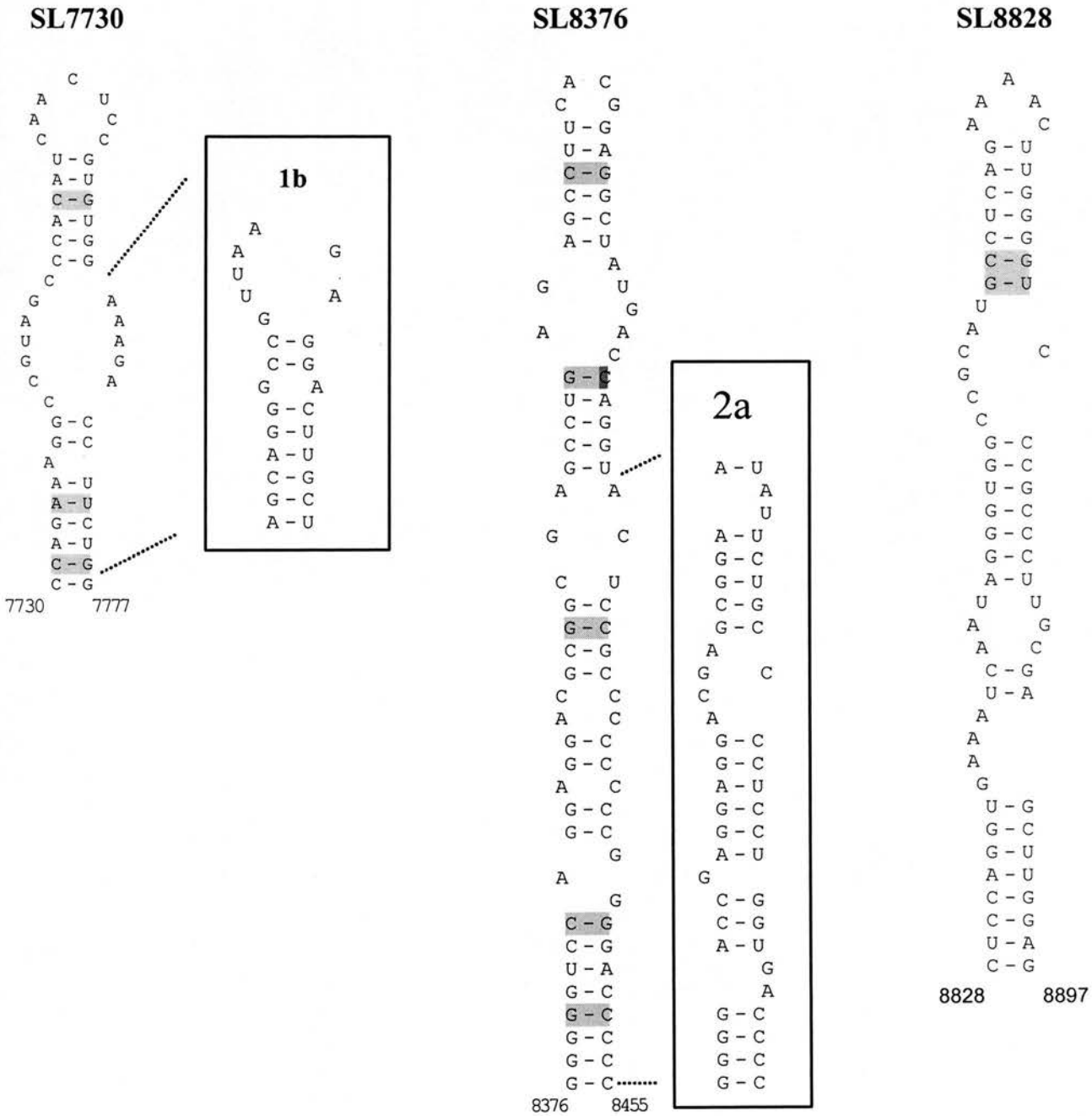
substitutions were directly adjacent to conserved single stranded bulges. However, in no instances were substitutions observed to destabilise the RNA structures in such a way as to cause slippage in nucleotide pairing or disrupt the overall conformation; the majority of nucleotide changes at synonymous sites either stabilised the stem loops or had a neutral effect (Table 4.1).

Predicted HCV NS5B RNA structures

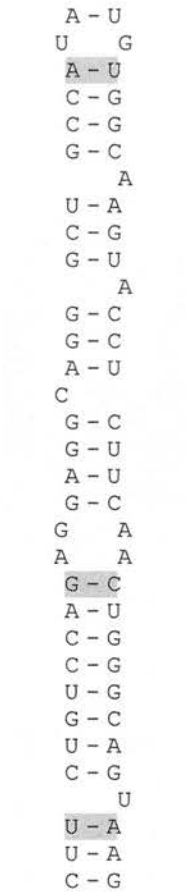
RNA structures **SL7730** and **SL8376** were the most upstream of all the HCV stem loops predicted within the NS5B region (Fig. 4.2). SL7730 was highly conserved in all HCV genotypes exhibiting two covariant substitutions (Table 4.2). SL7730 also included the greatest proportion of stem loop stabilising or neutral synonymous mutations. SL8376 was highly conserved in all genotype 1a, 1b and 2a sequences examined. In genotypes 3a, 4a, 5a and 6a the size and shape of the RNA structure was maintained even though there were a number of differences in predicted base pairings and the degree of sequence conservation. Four covariant sites were observed within SL8376.

RNA structure **SL8828** was highly conserved within genotypes 1b to 5a and nine out of ten 1a sequences examined (Table 4.2). In a single 1a sequence (D50409) and one out of six 6a sequences (JK049E1) examined there were a number of differences in the paired and unpaired regions, although the overall shape and size of the stem loops was maintained (Fig. 4.2). In the remaining 6a sequences examined, the overall size and shape of the RNA structure was not maintained but a structure of

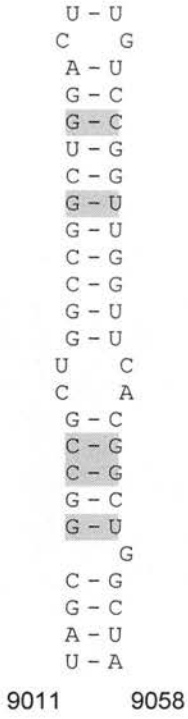
Figure 4.2. Schematic representation of HCV NS5B region conserved RNA stem loops. Covariant sites are highlighted in grey.



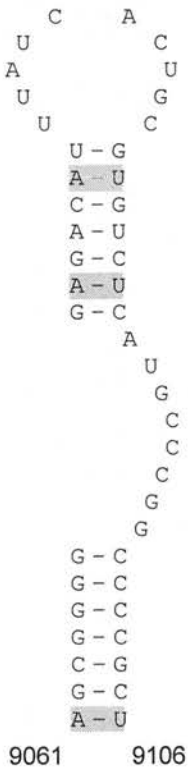
SL8926



SL9011



SL9061



SL9118

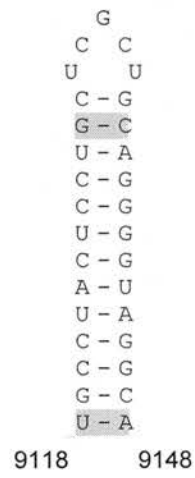


Figure 4.2: Schematic representation of HCV core gene and NS5B coding region conserved RNA stem loops. Stem loop names are given above the structures; positions of the upstream and downstream extent of base pairings are given below the structures and are numbered from the start of the HCV alignment (see appendix). Covariant sites are indicated by shading of paired nucleotides. Symbol '-' canonical Watson-Crick base pairing or G-U pairing. Examples of structurally different stem-loops (SL7730 and SL8376) shown in boxes (genotypes of variant structures indicated in bold).

different conformation was present within the same region. Nucleotide substitutions resulted in slippage of base pairings between all genotypes towards the base of the stem loop, which in most cases was resolved by nucleotide pair 8857-8877. Two covariant sites were observed within SL8828 (excluding genotype 6a which was too divergent to make a comparison).

SL8926 was the least highly conserved RNA structure predicted within HCV (Table 4.2). It was highly maintained in all genotype 1b and 2a sequences and six out of ten 1a sequences (AF011751, AF511948, AF11950, HEC278830, HPCHJ1 and HPCPLYPRE). In all other genotype sequences examined, single nucleotide substitutions towards the base of the structure led to slippage in base pairing and the formation of novel single stranded bulges. However, the overall shape, size and position of the stem loop was maintained in all subtypes and two covariant sites were observed (Fig. 4.2).

SL9011, **SL9061** and **SL9118** were the most down-stream RNA structures predicted within the coding region of HCV. SL9011 was directly adjacent to SL9061 which in turn was only twelve nucleotides upstream of the 5' nucleotide of SL9118 (Fig. 4.2). The 3' nucleotide of SL9118 was nineteen nucleotides upstream of the stop codon. All three RNA structures were highly conserved between genotypes 1a-6a (Table 4.2). Six covariant sites were observed within SL9011, three within SL9061 and one within SL9118. A short slippage in base pairing, within and between genotypes, was observed between the four nucleotide pairs at the base of SL9011 which was resolved by nucleotide pair five (9015-9054). Sequence variation between positions 9015 and 9054 also led to the formation of unique single stranded bugles within the main paired stem of SL9011 in genotype 1a but did not result in

further slippage in nucleotide pairing. No further regions of slippage were observed for RNA structures SL9011, SL9061 and SL9118. Of note was an eight nucleotide unpaired bulge in SL9061 between positions 9092 and 9099; the sequence within this unpaired region included three synonymous sites all of which are very highly conserved.

4.2.2 HGV/GBV-C SECONDARY STRUCTURE PREDICTION

NS5B encoding region

Eight thermodynamically stable and evolutionarily conserved RNA structures were predicted within the NS5B region of HGV/GBV-C (Fig. 4.3). The RNA structures were provisionally named according to their position in relation to the other stem loops and the stop codon. Structures $SL_{NS5B}VIII$, $SL_{NS5B}VII$, $SL_{NS5B}VI$, $SL_{NS5B}V$, $SL_{NS5B}IV$, $SL_{NS5B}III$, $SL_{NS5B}II$ and $SL_{NS5B}I$ spanned regions 8250 to 8302; 8309 to 8321; 8327 to 8395; 8402 to 8500; 8551 to 8585; 8640 to 8730; 8761 to 8789 and 8802 to 8821 respectively.

All the predicted RNA structures exhibited a high level of conservation between genotypes and in most cases with HGV/GBV-C_{cpz} (Table 4.3). However, variation in structure length, sequence and paired nucleotides was observed within a number of predicted RNA structures. Covariant sites were less frequent within HGV/GBV-C predicted RNA structures than was the case within HCV. This may have been due to the higher levels of nucleotide homology within the regions of HGV/GBV-C

Figure 4.3: Schematic representation of HGV/GBV-C NS5B coding region conserved RNA stem loops. Covariant sites are indicated by grey boxes.

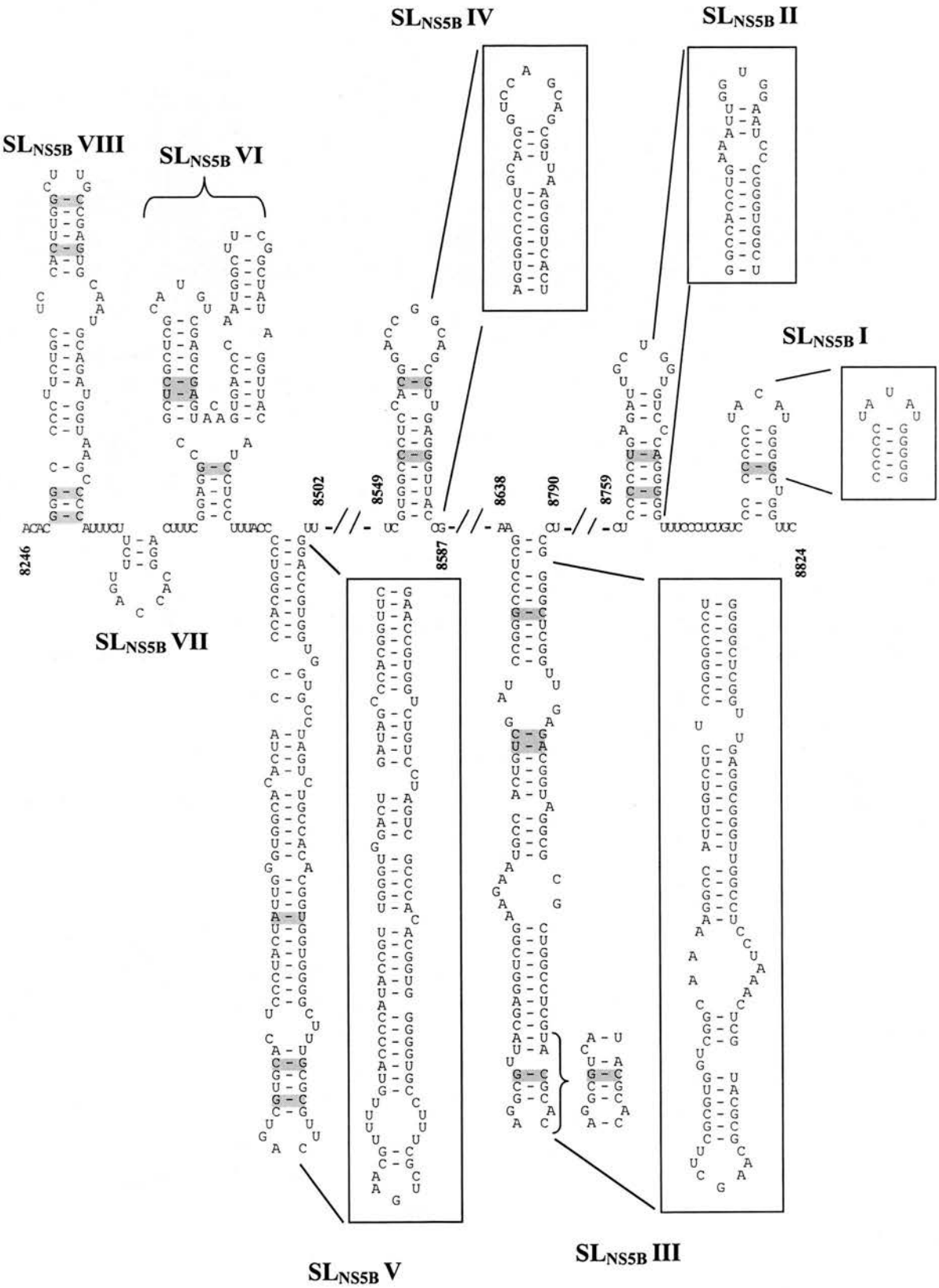


Figure 4.3 legend. Schematic representation of HGV/GBV-C conserved RNA structures within the NS5B encoding region and 3'UTR. RNA structure names are given below and above the stem loops. Bases are numbered from the start of the HGV/GBV-C alignment (see appendix). Symbols: '-' canonical Watson-Crick base pairing or G-U pairing; '//', intervening sequence without secondary structure. Covariant sites are indicated by shading of paired nucleotides. HGV/GBV-C_{CPZ} RNA structures are shown in boxes.

Table 4.3 Structure conservation of predicted stem loops within the NS5B coding region of HGV/GBV-C

Stem-loop	genotype	N[u]*	+++ %	++ %	+ %	Co-variance	Stable variance†
NS5B VIII	1	3	100	0	0	0	5/8
	2	9	100	0	0	0	5/6
	3	8	100	0	0	0	7/11
	4	3	100	0	0	0	4/5
	CPZ	1	0	0	0	/	/
NS5B VII	1	3	100	0	0	0	1/1
	2	9	100	0	0	0	2/2
	3	8	100	0	0	0	3/3
	4	3	100	0	0	0	2/2
	CPZ	1	0	0	0	/	/
NS5B VI	1	3	100	0	0	1	6/6
	2	9	100	0	0	1	7/7
	3	8	100	0	0	0	9/13
	4	3	33.3	66.7	0	0	7/11
	CPZ	1	0	0	100	/	/
NS5B V	1	3	66.7	33.3	0	0	7/9
	2	9	100	0	0	0	10/13
	3	8	100	0	0	0	6/9
	4	3	100	0	0	0	7/10
	CPZ	1	0	100	0	/	/
NS5B IV	1	3	100	0	0	1	2/2
	2	9	100	0	0	1	1/1
	3	8	100	0	0	1	0
	4	3	100	0	0	0	0
	CPZ	1	100	0	0	0	0
NS5B III	1	3	100	0	0	3	3/5
	2	9	100	0	0	3	2/4
	3	8	100	0	0	3	3/6
	4	3	100	0	0	1	1/1
	CPZ	1	0	100	0	/	/
NS5B II	1	3	100	0	0	0	1/2
	2	9	100	0	0	1	0/2
	3	8	100	0	0	2	0/1
	4	3	100	0	0	2	1/3
	CPZ	1	100	0	0	/	/
NS5B I	1	3	100	0	0	0	1/1
	2	9	100	0	0	1	1/1
	3	8	100	0	0	1	1/1
	4	3	100	0	0	1	1/1
	CPZ	1	100	0	0	/	/

Table 4.3: Structures were scored from + to +++ depending on the degree of similarity to the most common structure. +++ Stem-loop structurally identical; ++ minor differences in base pairing and/or size but the overall stem-loop conservation was maintained; + different structure in the same region. The frequency within which single synonymous substitutions retain or had a neutral effect on stem-loop structure is also shown. * Number of sequences examined for a given genotype: † Number of stabilising or neutral synonymous substitutions compared to the total number of synonymous substitution (semi-covariance).

sequence analysed as compared to HCV. However, covariant sites were observed in seven of the eight RNA structures predicted within the NS5B encoding region of HGV/GBV-C (SL_{NS5B}VII which only has three paired nucleotides was the exception). As with HCV, the third base codon positions usually paired with down-stream third base positions and synonymous covariant sites were observed within five of the seven stem loops with covariant substitutions (SL_{NS5B}VIII and SL_{NS5B}III being the exception). Non synonymous covariant sites were observed within SL_{NS5B}VIII, SL_{NS5B}VI, SL_{NS5B}V and SL_{NS5B}III. All the predicted RNA structures exhibited a large excess of synonymous substitutions which either stabilised or had a neutral effect on stem loop structure.

SL_{NS5B}VIII and **SL_{NS5B}VII** were most upstream of all the RNA structures predicted within the NS5B region of HGV/GBV-C and were very highly conserved between all the human isolate sequences examined (Table 4.3). SL_{NS5B}VII was the smallest independent RNA structure predicted within HGV/GBV-C with only three paired nucleotides and no observed covariant substitutions (Fig. 4.3). Four covariant sites were observed within SL_{NS5B}VIII. No comparable structures were observed within HGV/GBV-C_{CPZ}.

SL_{NS5B}VI was made up of two nucleotide pair stacks branching off a basal structure which was composed of five nucleotide pairs (Fig. 4.3). This "Y shaped" three-way junction was unique amongst all the stem loops predicted within this study for both HGV/GBV-C and HCV. SL_{NS5B}VI was conserved amongst all genotypes with the exception of two out of three genotype 4 sequences examined, in which the structure was not branched (AB018667 and AB021287). Three covariant substitutions were observed; one within the basal structure and two within the

downstream branch. $SL_{NS5B}VI$ was not conserved within HGV/GBV-C_{CPZ}, although two thermodynamically stable structures were observed within the same region between nucleotides 8313 to 8349 and 8353 and 8392.

$SL_{NS5B}V$ and **$SL_{NS5B}III$** were the longest RNA structures predicted in HGV/GBV-C. $SL_{NS5B}V$ was thirty eight nucleotide pairs in length and $SL_{NS5B}III$ thirty four (Fig. 4.3). Both RNA structures were highly conserved within all HGV/GBV-C genotypes (Table 4.3). A UUC codon insertion was observed at position 8458 within $SL_{NS5B}V$ in one of three genotype 1 sequences examined (U36380). This insertion was in frame and within the single stranded pre-terminal bulge; the overall nucleotide pairing of the stem loop was unaffected. Three covariant substitutions were observed within $SL_{NS5B}V$ and four within $SL_{NS5B}III$. The position and length of each stem loop was conserved within HGV/GBV-C_{CPZ} but in both cases there was slippage in nucleotide pairings due to sequence variation.

$SL_{NS5B}IV$, **$SL_{NS5B}II$** and **$SL_{NS5B}I$** were highly conserved between all HGV/GBV-C genotypes and HGV/GBV-C_{CPZ}. Two covariant substitutions were observed within $SL_{NS5B}IV$ which was flanked by the much longer RNA structures $SL_{NS5B}V$ and $SL_{NS5B}III$ (Table 4.3). Two covariant substitutions were also observed within $SL_{NS5B}II$ and one within $SL_{NS5B}I$ which were the most down-stream of all the predicted NS5B RNA structures. Both $SL_{NS5B}IV$ and $SL_{NS5B}II$ were one nucleotide pair longer within HGV/GBV-C_{CPZ} due to pairing of nucleotides 8550-8586 and 8760-8790 respectively at the base of each stem loop within the chimpanzee isolate (Fig. 4.3). $SL_{NS5B}I$ was two nucleotide pairs shorter in HGV/GBV-C_{CPZ} due to a single G to A substitution at position 8820 which disrupted the two nucleotide pairs at the base of the stem-loop.

4.2.3 HGV/GBV-C 3'UTR SECONDARY STRUCTURE

Seven thermodynamically stable evolutionarily conserved RNA structures were predicted within the 3'UTR of HGV/GBV-C (Fig. 4.4). The RNA structures were provisionally named according to their position in relation to the other stem loops and the stop codon. Structures 3'SLVII, 3'SLVI, 3'SLV, 3'SLIV, 3'SLIII, 3'SLII and 3'SLI spanned regions 8936 to 8958; 8959 to 8975; 8987 to 9025; 9074 to 9087; 9088 to 9101; 9170 to 9195 and 9199 to 9241 respectively.

The RNA structures predicted within the 3'UTR were more highly conserved between genotypes than those of the NS5B encoding region (Table 4.4). However, the extreme downstream region of the 3'UTR was incomplete or missing in a number of sequences, resulting in a smaller sample size being available for examination. Alignment with homologous HGV/GBV-C sequences also indicated that the terminal 55 nucleotides of HGV/GBV-C_{CPZ} were missing, including the regions forming 3'SLII and 3'SLI within the human isolate sequences. Due to the high level of sequence conservation within the 3'UTR of HGV/GBV-C the levels of stabilising, de-stabilising and structurally neutral substitutions were much lower than within the NS5B coding region.

3'SLVII and **3'SLVI** are the most upstream of all the predicted 3'UTR RNA structures. The 5' nucleotide at the base of 3'SLVII is twenty five nucleotides down-stream of the stop codon and the 3' nucleotide is directly adjacent to 3'SLVI (Fig. 4.4). 3'SLVII is highly conserved between all genotypes and six covariant sites were observed out of eight nucleotide pairs (Table 4.4). 3'SLVII was not present within

Figure 4.4: Schematic representation of HGV/GBV-C 3'UTR conserved RNA stem loops. Covariant sites are indicated by grey boxes.

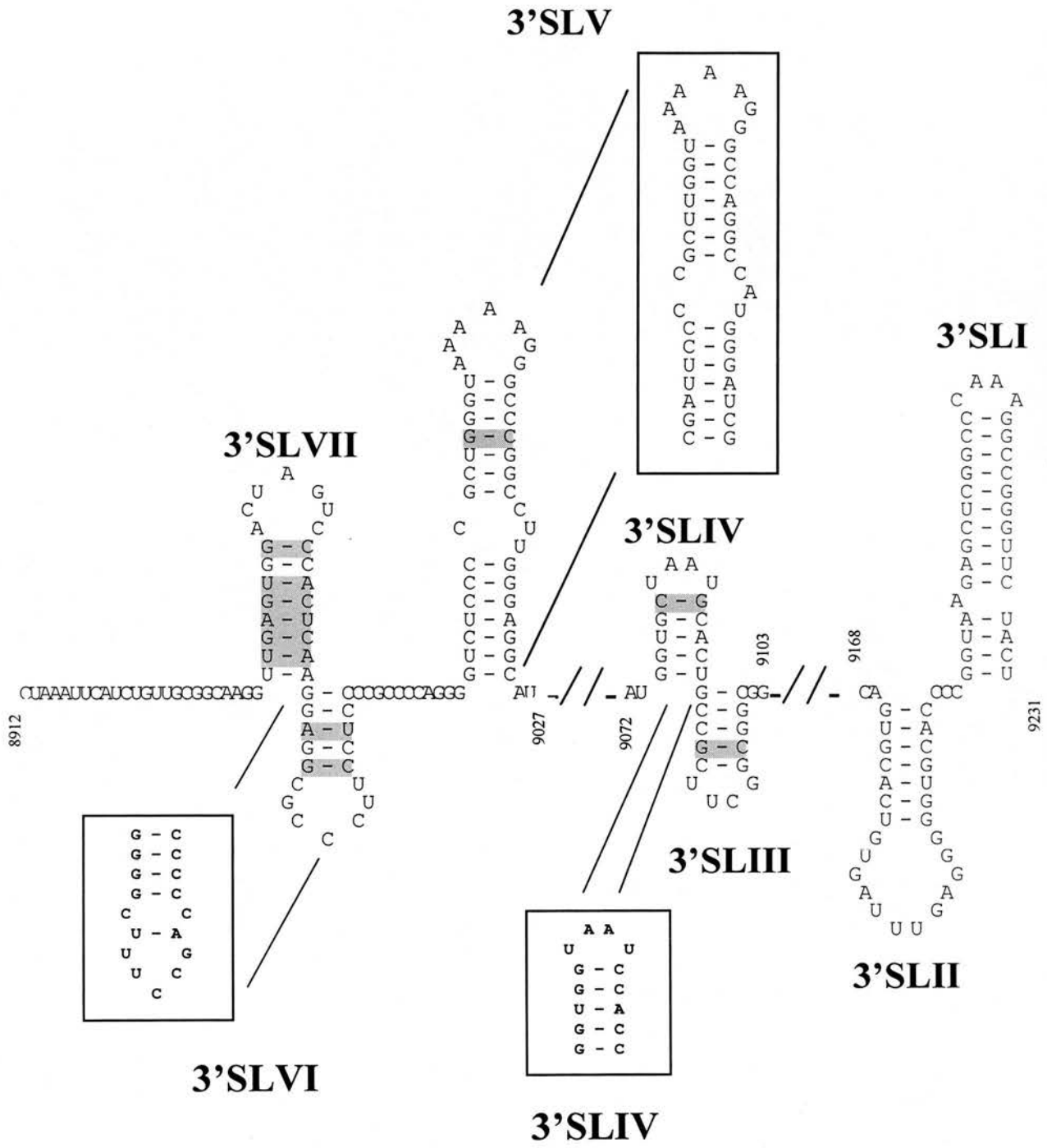


Figure 4.4 legend. Schematic representation of HGV/GBV-C conserved RNA structures within the NS5B encoding region and 3'UTR. RNA structure names are given below and above the stem loops. Bases are numbered from the start of the HGV/GBV-C alignment (see appendix). Symbols: '-' canonical Watson-Crick base pairing or G-U pairing; '//', intervening sequence without secondary structure. Covariant sites are indicated by shading of paired nucleotides. HGV/GBV-C_{CPZ} RNA structures are shown in boxes.

Table 4.4 Structure conservation of predicted stem loops within the 3' UTR of HGV/GBV-C

Stem-loop	genotype	N[u]*	+++ %	++ %	+ %	Co-variance	Stable variance†
3'SL VII	1	2	100	0	0	1	0/0
	2	7	100	0	0	4	0/0
	3	4	100	0	0	2	0/2
	4	2	100	0	0	2	0
	CPZ	1	0	0	0	/	/
3'SL VI	1	2	100	0	0	0	0/0
	2	7	100	0	0	0	0/0
	3	4	75	0	25	0	1/1
	4	2	100	0	0	0	0/0
	CPZ	1	100	0	0	/	/
3'SL V	1	2	100	0	0	0	0/0
	2	6	100	0	0	0	0/0
	3	4	100	0	0	0	0/0
	4	2	100	0	0	0	1/1
	CPZ	1	100	0	0	/	/
3'SL IV	1	2	100	0	0	0	0/0
	2	6	100	0	0	0	0/0
	3	4	100	0	0	0	0/0
	4	2	100	0	0	0	1/1
	CPZ	1	100	0	0	/	/
3'SL III	1	2	100	0	0	0	0/0
	2	6	83.3	16.7	0	1	1/1
	3	3	0	0	100	/	/
	4	2	100	0	0	0	0/0
	CPZ	1	0	0	0	/	/
3'SL II	1	2	100	0	0	0	0/0
	2	5	100	0	0	0	0/0
	3	3	100	0	0	0	0/0
	4	2	100	0	0	0	2/3
	CPZ	0	/	/	/	/	/
3'SL I	1	2	100	0	0	0	0/0
	2	3	100	0	0	0	1/1
	3	3	100	0	0	0	0/0
	4	2	100	0	0	0	2/3
	CPZ	0	/	/	/	/	/

Table 4.4: Structures were scored from + to +++ depending on the degree of similarity to the most common structure. +++ Stem-loop structurally identical; ++ minor differences in base pairing and/or size but the overall stem-loop conservation was maintained; + different structure in the same region. The frequency within which single synonymous substitutions retain or had a neutral effect on stem-loop structure is also shown. * Number of sequences examined for a given genotype: † Number of stabilising or neutral synonymous substitutions compared to the total number of synonymous substitution (semi-covariance).

HGV/GBV-C_{CPZ} due to a thirteen nucleotide deletion between positions 8946 to 8958 which overlapped with the RNA structure. Two covariant sites were observed within 3'SLVI which was highly conserved between genotypes 1, 2, 4, HGV/GBV-C_{CPZ} and three out of the four genotype 3 sequences examined. In one of four genotype 3 sequences examined (AB008342) a tri-nucleotide deletion between positions 8961 to 8964 resulted in slippage of base pairing.

3'SLV was highly conserved between all genotypes of HGV/GBV-C and HGV/GBV-C_{CPZ} (Table 4.4). No covariant sites were noted within the human isolate sequences, although one covariant site was observed between the HGV/GBV-C_{CPZ} and HGV/GBV-C RNA structures. Nucleotide substitutions in both the 5' and 3' extremes of 3'SLV within HGV/GBV-C_{CPZ} were predicted to result in slippage of nucleotide pairings within the chimpanzee isolate. This was restricted to the basal nucleotide pair stack and was resolved within the single stranded internal bulge at nucleotide 8994.

3'SLIV and **3'SLIII** were both short RNA structures of only five predicted nucleotide pairs each (Fig. 4.4). The terminal downstream nucleotide of 3'SLIV was directly adjacent to the first upstream nucleotide of 3'SLIII. 3'SLIV was highly conserved between all human genotypes and HGV/GBV-C_{CPZ} with one covariant site observed between the human and chimpanzee isolates (Table 4.4). 3'SLIII was highly conserved in all genotype 1 and 4 sequences and five of six genotype 2 sequences examined. In a single genotype 2 sequence (AB013501) the terminal C-G nucleotide pair, at positions 9043 and 9048, was absent. One covariant site was observed within the structure between genotypes 1, 2 and 4. 3'SLIII was absent from all the genotype 3 sequences examined due to the deletion of nucleotides 9043

and 9044. The stem-loop was also absent from HGV/GBV-C_{CPZ} due to the deletion of nucleotides 9044 and 9055.

3'SLII and **3'SLI** were the most down-stream of all the predicted RNA structures within the 3'UTR of HGV/GBV-C. 3'SLI formed a terminal stem loop of thirteen paired nucleotides which incorporated the extreme 3' end of the HGV/GBV-C sequence (Fig. 4.4). The 3' terminal nucleotide of 3'SLII (position 9195) is separated from the 5' nucleotide of 3'SLI (position 9199) by a CCC unpaired triplet. 3'SLII is a shorter RNA structure with seven paired nucleotides and a long unpaired terminal loop between positions 9177 and 9188. Both RNA structures were conserved between all genotypes; unfortunately it was not possible to make a comparison with HGV/GBV-C_{CPZ} as the terminal 55 nucleotides were missing from this sequence (Table 4.4).

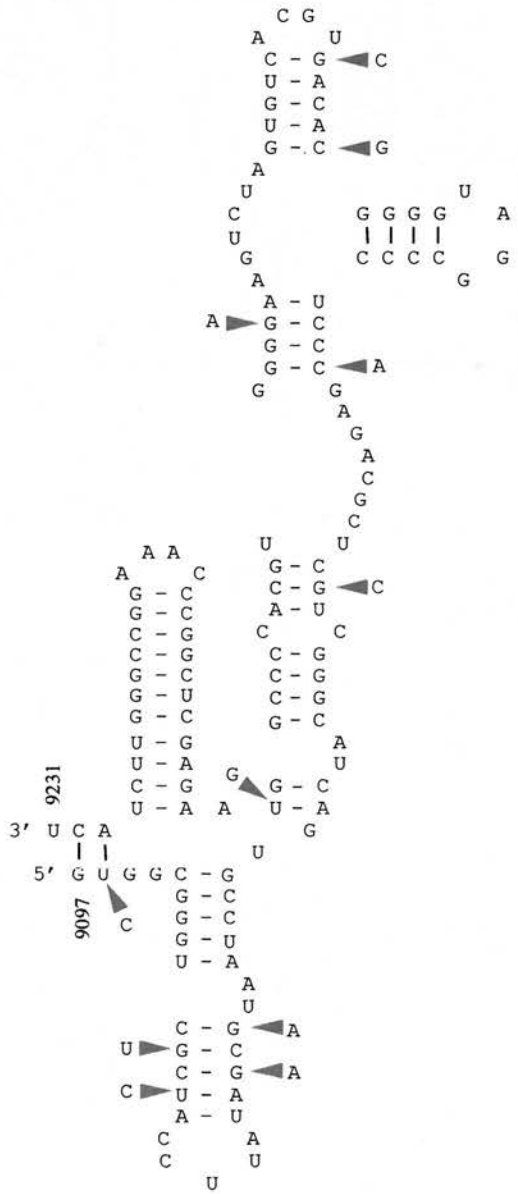
4.3 DISCUSSION

In this study free energy minimisation on folding, structure conservation between genotypes, the occurrence of covariant substitutions and stabilising or neutral synonymous substitution in nucleotide paired regions have been used to predict and map evolutionarily conserved and thermodynamically stable RNA structures within both HCV and HGV/GBV-C. Specific RNA structure predictions were made for the 3'UTR of HGV/GBV-C (Fig. 4.4), in which excess free energy on folding was observed (chapter 3) (Simmonds and Smith, 1999b; Cuceanu et al., 2001). Structural predictions were also made for the NS5B regions of both HCV and HGV/GBV-C (Figs. 4.2 and 4.3), and the core gene of HCV (Fig. 4.1), in which reductions in synonymous variability and clustering of covariant substitutions were observed as well as excess free energy on folding (chapter 3) (Simmonds and Smith, 199b; Cuceanu et al., 2001; Tuplin et al., 2002).

4.3.1 RNA STRUCTURE OF THE HGV/GBV-C 3'UTR

Predictions made in this study of seven thermodynamically stable evolutionarily conserved RNA structures within the 3'UTR of HGV/GBV-C differ from previous studies which analysed a very limited number of sequences and concentrated on only the extreme downstream region of the 3'UTR. A previous model, in which only the terminal 140 nucleotides of three sequences were analysed, proposed four terminal RNA structures within the downstream region of the 3'UTR, (Fig. 4.5) (Okamoto et

Figure 4.5: Schematic model of the RNA structure for the terminal 140 nucleotides of the HGV/GBV-C 3'UTR, proposed by Okamoto et al (1997). Substitutions which disrupt the proposed nucleotide pairings are indicated with a red triangle '◄'.



al., 1997). Analysis of the forty two sequences used in this study identified eleven substitutions which disrupt the three downstream RNA structures predicted in this model. For example, the proposed U-A pairing between nucleotides 9098 and 9229 is disrupted by a U to C substitution at position 9229 in genotypes 1, 2 and 3; the proposed C-G nucleotide pairing between positions 9150 and 9189 is disrupted by a C to G substitution in genotype 3 and a C to A substitution in genotype 4, at position 9150. Substitutions also disrupt proposed nucleotide pairings C-G between positions 9106 and 9120; G-C between positions 9107 and 9119; C-G between positions 9108 and 9118; U-A between positions 9109 and 9117; A-U between positions 9130 and 9201; G-C between positions 9140 and 9193; C-G between positions 9152 and 9187; C-G between positions 9166 and 9179 and G-C between positions 9170 and 9175. In addition the Okamoto model is not supported by covariant substitutions and is limited to the terminal 140 nucleotides of the 3'UTR.

The model of 3'UTR RNA structure proposed by Okamoto and others predicts a terminal stem loop with eleven nucleotide pairs between positions 9203 and 9228 (Fig. 4.5). A separate study in which twelve 3'UTR sequences were analysed predicted the same stem loop at the terminal region of the 3'UTR but failed to predict any further upstream structures (Katayama et al., 1998). Again the region of sequence analysed was restricted to the extreme downstream region the 3'UTR. This terminal RNA structure shows a high degree of homology with 3'SLI predicted in this study. However, the final three nucleotides of the 3'UTR, which are predicted to form the terminal three paired nucleotides of 3'SLI in this study, are shown as unpaired in the previous models.

Highly conserved RNA structures have been predicted by free energy minimisation, phylogenetic and RNase cleavage analysis in the terminal down-stream regions of a number of different viruses including pestiviruses such as bovine viral diarrhea virus (BVDV) (Yu et al., 1999; Deng and Brock, 1993), the *Picornaviridae* family (Witwer et al., 2001; Mellits et al., 1998), most flaviviruses (Proutski et al., 1997; Brinton et al., 1986; Rice et al., 1985; Rauscher et al., 1997; Olsthoorn and Bol, 2001), GB viruses A (GBV-A) and B (GBV-B) (Sbardellati et al., 1999) and HCV (Kolykhalov et al., 1996; Tanaka et al., 1996; Blight and Rice, 1997).

The terminal 3'UTR sequences of HCV, GBV-B and GBV-A are all predicted to form three similarly structured adjacent stem-loops in a "cloverleaf" conformation. The configuration of 3'SLI and 3'SLII of GBV-C closely resembles the two most downstream RNA structures predicted for these related viruses, however there is no evidence for a third adjacent upstream stem loop similar to that of HCV, GBV-A or GBV-B (fig. 4.4). The predicted terminal stem loops of HCV and GBV-B are also longer (19 and 21 nucleotide pairs respectively) than the HGV/GBV-C 3'SLI homologue (14 nucleotide pairs) which is closer in size to the equivalent predicted GBV-A stem-loop (11 nucleotide pairs). 3'SLII in HGV/GBV-C (7 nucleotide pairs) more closely resembles the predicted HCV homologue (8 nucleotide pairs) than those of either GBV-A or GBV-B (4 and 3 nucleotide pairs respectively).

The high degree of 3'UTR sequence conservation which is observed between human HGV/GBV-C genotypes, with mean pairwise distances of between 3.87% to 6.6%, contrasts with that between human and chimpanzee isolates of between 12.8% to 13.4% (Cuceanu et al., 2001). However, despite a relatively high level of

sequence divergence 3'SLVI, 3'SLV and 3'SLIV were all predicted within HGV/GBV-C_{CPZ}, with a high level of structural conservation (Table 4.4). It was not possible to make structural predictions for the most downstream region of the HGV/GBV-C_{CPZ} 3'UTR as it was shown, after alignment with human genotypes, that the terminal 55 nucleotides were missing from the sequence.

The 3'UTR of HCV is absolutely required for infectivity in the chimpanzee animal model (Yanagi et al., 1999). It has also been shown to bind polypyrimidine tract-binding protein (PTB) and NS3 (Banerjee and Dasgupta, 2001; Ito and Lai, 1997), suggesting a role in viral replication and/or regulation of translation. Apart from this it remains unclear what role RNA structure within the 3'UTR may play, although the level of 3'UTR secondary structure homology between a range of virus types suggests a high degree of functional importance. (Possible functional roles are discussed in more detail in chapter 5).

4.3.2 RNA STRUCTURE WITHIN THE POLYPROTEIN CODING REGIONS OF HCV AND HGV/GBV-C

The important role played by RNA stem loop structures within the polyprotein coding regions of RNA viruses is becoming increasingly recognised. For example, small stem loops within the coding regions of rhinovirus type 14 (McKnight and Lemon, 1998), enteroviruses (Goodfellow et al., 2000) and cardioviruses (Lobert et al., 1999) have all been shown to play an important role in virus replication. In this study thermodynamic and phylogenetic methods have been used to predict three

RNA structures within the core gene (Fig. 4.1), seven within the NS5B regions of HCV (Fig. 4.2) and eight within the NS5B encoding region of HGV/GBV-C (Fig. 4.3) (Tuplin et al., 2002; Cuceanu et al., 2001).

A previously published model predicted two conserved RNA structures within the core gene (between positions 47 to 84, 87 to 167) and two within the NS5B region (between positions 9014 to 9054 and 9118 to 9148) of HCV (Smith and Simmonds, 1997a). The core gene stem loops predicted in this model overlap the same regions and exhibit very high levels of structural homology to SL47 and SL87. In the Smith and Simmonds model a total of twenty six epidemiologically unlinked sequences were analysed, including sixteen 1b genotype sequences, three 1a and two 3a sequences and single examples of 2a and 4a sequences; no 5a or 6a sequences were analysed. In the present study a total of one hundred and twenty one epidemiologically unlinked sequences were analysed, including both genotype 5a and 6a sequences. This increased data set enabled a more accurate estimate of RNA structure conservation to be made.

SL47 exhibits almost complete structural homology to the Smith and Simmonds model. However, nucleotide pair C-G (between positions 64 and 70), which is adjacent to the terminal single stranded loop of SL47 is unpaired in the previous model due to a G to U substitution at position 70. This substitution was not observed in the current study.

The initial six and terminal fourteen nucleotide pairs of SL87 are identical to the previous model. However, nucleotide pair U-A (positions 93 and 161) was shown as unpaired and nucleotides 97 (G) and 159 (C) as paired in the Smith and Simmonds model which contradicts the current study. In a number of genotypes (1a, 1b, 2a, 4a

and 5a) nucleotide substitutions disrupt the previously predicted pairing between positions 97 and 159. Nucleotides 104 (C) and 150 (G) are also predicted to pair in the previous model but not in the current study. However, nucleotide substitutions in genotypes 1a, 3a, 4a, 5a and 6a would again disrupt this previously predicted pairing. These contradictory predictions result in a slippage in nucleotide pairings between base pairs 87-67 (U-A) and 107-146 (G-C) in the previous model and would disrupt a covariant substitution, between nucleotides 107 and 146, observed in the current study.

RNA structures were also predicted in the Smith and Simmonds model between positions 264 to 310, 337 to 394 and 443 to 475, although the structures of these potential stem loops were not mapped. The latter region overlaps SL443 which is the most downstream of all the in the core gene RNA structures predicted in HCV. No evolutionarily conserved stem loops or covariant substitutions were observed between regions 264 to 310 or 337 to 394. However, it is possible that RNA structures exist within these regions whose function is not dependent upon a conserved structural conformation. Such structures would be difficult to predict and map using phylogenetic conservation methods and would require mutational analysis of an infectious clone in a chimpanzee model, as the structural genes are missing from the more amenable replicon system, in order to assess functionality.

A further study, in which eight sequences were analysed, has previously predicted two evolutionarily conserved thermodynamically stable RNA structures within the downsteam sequence of the NS5B region of HCV, between nucleotides 9015 to 9054 and 9118 to 9148 (Han and Houghton, 1992). The two stem loops predicted in the Han and Houghton study are structurally identical with the two NS5B stem loops

predicted in the Smith and Simmonds study. The most downstream of these RNA structures is also identical to SL9118. The upstream structure predicted in the previous models shows a degree of homology to SL9011, although the base of the structure is shown to be formed by a C-G pair between positions 9054 and 9118. This differs from the current model in which a U-A pair between positions 9011 and 9058 forms the base of the stem loop. This discrepancy in the structural prediction of SL9011 is due to the prediction in the previous models that nucleotides 9014 (C) and 9054 (G) form a pair at the base of the RNA structure. However, a nucleotide substitution in genotype 1 at position 9014 (G to A) would disrupt this pairing. In the current study nucleotide 9014 (C) is predicted to pair with 9055 (G) and nucleotide 9054 (G) is unpaired, resulting in an extra three nucleotide pairs at the base of the structure compared to that of the previous model. Neither the Han and Houghton or the Smith and Simmonds models predict SL9061 and show the region between SL9017 and SL9124 as unpaired. The discrepancies between the conformation and presence of RNA structures predicted between the different models may be due to the limited number of sequences previously available and the short distance over which the sequences were analysed.

An excess free energy on folding and reduction in variability at synonymous sites has been observed towards the 5' and 3' ends of the polyprotein reading frames of both HCV (Smith and Simmonds, 1997a; Tuplin et al., 2002) (chapter 3) and HGV/GBV-C (Simmonds and Smith, 1999b) (Cuceanu et al., 2001) (chapter 3). These locations correspond to the core gene and NS5B coding regions of HCV and the E1/E2 genes and NS5B coding region of HGV/GBV-C. Four RNA structures have previously been predicted, by analysis of covariant substitutions, within the E1

and E2 genes of HGV/GBV-C and one between positions 265 to 309 was structurally mapped. This RNA structure was observed to contain three covariant substitutions and is conserved between HGV/GBV-C, HGV/GBV-C_{CPZ} and GBV-A (Simmonds and Smith, 1999). Given that these RNA structures and those predicted in this study are evolutionarily conserved and thermodynamically stable it is likely that they account for the bias in nucleotide substitutions and excess free energy on folding observed towards the extreme ends of the polyprotein coding regions of both HCV and HGV/GBV-C (chapter 3).

To a lesser extent free energy and nucleotide substitution biases are observed within other regions of the HGV/GBV-C polyprotein coding regions. For example, an excess free energy on folding and a reduction in synonymous variability occurs between positions 6001 to 7000 (NS5A encoding region) and 4001 to 5000 (NS3 encoding region) (Simmonds and Smith, 1999) (Cuceanu et al., 2001). Although secondary structures were not mapped within these regions, it seems likely that they will contain RNA structures. Indeed the excess free energy on folding across the genome of HGV/GBV-C is much higher then for HCV, in which it is more restricted to the 5' and 3' extremes, suggesting further RNA structures throughout the genome.

CHAPTER 5

VISUALISATION OF RNA STRUCTURE

5.1 INTRODUCTION

Fifteen evolutionary conserved RNA structures were previously predicted within the genome of HGV/GBV-C, using both thermodynamic and phylogenetic methods (chapters 3 and 4) (Cuceanu et al., 2001). Eight stem loops were predicted and mapped within the NS5B region at the 3' extreme of the polyprotein coding region and seven within the 3'UTR. Despite sequence divergence each of the predicted RNA structures were shown to be highly conserved between all genotypes and multiple covariant and semi-covariant substitutions were observed suggesting functional constraint. However, the free energy minimisation algorithm used previously does not predict higher order folding (tertiary structure), such as pseudoknots. In order to physically investigate the position and tertiary folding structure of the predicted stem loops, RNA transcripts of both the NS5B region and 3'UTR of HGV/GBV-C, were directly visualised under transmission electron microscopy in chapter 5.

In chapter 3 it was shown that excess FFE is not confined to specific regions of the HGV/GBV-C sequence but that the complete genome itself may be highly structured (Cuceanu et al., 2001). This is consistent with phylogenetic results which show that reductions in synonymous variability and covariant substitutions also occur along the complete length of the ORF (Simmonds and Smith, 1999b; Cuceanu et al., 2001). In order to investigate the potential of genome wide RNA folding, complete RNA transcripts of HGV/GBV-C were also visualised by electron microscopy in chapter 5.

SEQUENCES ANALYSED

Subgenomic RNA transcripts were transcribed *in vitro* using cloned cDNA NS5B region and 3'UTR as template. The subgenomic fragments were originally amplified by RT-PCR from HGV/GBV-C positive serum (genotype 1), which had been collected from a haemophilia patient. A full length infectious cDNA clone was used as a temple for the generation of genome length RNA transcripts (genotype 1) (GenBank Ac. No. AB013500).

5.2 RESULTS

In chapter 4 the secondary structure of eight stem loops in the NS5B region and seven in the 3'UTR were predicted using a combination of thermodynamic and phylogenetic methods. In chapter 3 the FFED results were consistent with genome wide RNA folding. In order to physically investigate the RNA folding structure of the NS5B region, 3'UTR and complete genome, RNA transcripts were negatively stained and visualised under a transmission electron microscope (section 2.4).

The downstream domain of the NS5B coding region (position 8410-8894) and 3'UTR (position 8875-9231) were reverse transcribed and then amplified by PCR from HGV/GBV-C positive human serum (genotype 1) (section 2.1); the sequences and genotypes of both regions were confirmed by cycle sequencing of each amplification product (section 2.2.7). The PCR products of both regions were then cloned into pGEM-T Easy vectors, containing both T7 and SP6 RNA transcription promoters at alternate ends of the vector ligation site (section 2.2). The ligation products were used to transform JM109 competent cells which were then cultured and plated out on LB selective plates. Insert orientation was assessed by single round PCR screening of individual bacterial colonies (section 2.2.5). Primers were designed to amplify across the insert boundary, with one primer designed to hybridise to the vector and the other to the cDNA insert itself. The orientation of individual inserts was then assessed through the presence or absence of an amplification product. This was performed across both the T7 (5') and SP6 (3') promoter boundaries so that each colony was checked twice for both a positive and

negative PCR result. Colonies in which the insert was shown to be in a sense orientation for the T7 RNA polymerase promoter were selected, cultured, and plasmid DNA extracted by midi-prep (section 2.2.6). The plasmid DNA was then linearised upstream of the cDNA insert, which was then used as a template from which RNA was transcribed *in vitro* using the T7 RNA polymerase promoter (section 2.3). Complete genome RNA transcripts were generated by transcription *in vitro* from a T7 RNA polymerase promoter, using a linearised full length clone as a template (genotype 1) (GenBank Ac. No. AB013500). RNA integrity was assessed by electrophoresis though a 5% denaturing acrylamide gel (representative examples are shown in Fig. 5.1).

After purification the RNA was slowly re-natured under physiological ionic conditions, hybridized with biotinylated oligonucleotides and labelled with streptavidin-coated colloidal gold particles (section 2.4). The RNA was then adsorbed onto a carbon film, negatively stained with a solution of uranyl acetate and mounted on gold plated copper grid. Micrographs of each specimen were then recorded on a transmission electron microscope.

5.2.1 VISUALISATION OF THE NS5B CODING REGION AND 3'UTR RNA FOLDING CONFORMATIONS

Visualisation of the NS5B coding region of HGV/GBV-C revealed a structure folded into a cruciform-like conformation with four major stems of similar length radiating from a central axis (Fig. 5.2). The four stems were arbitrarily labelled A,

Figure 5.1. Determination of HGV/GBV-C RNA transcript integrity and size after electrophoresis through a 5% denaturising acrylamide gel. Total RNA is stained with 0.04 % Methylene blue solution (section 2.3.5). **A:** NS5B region and 3'UTR (RNA transcripts from 3 independent clones shown for each region) **B:** Complete genome transcript (IC). Lane S in each case represents RNA size marker.

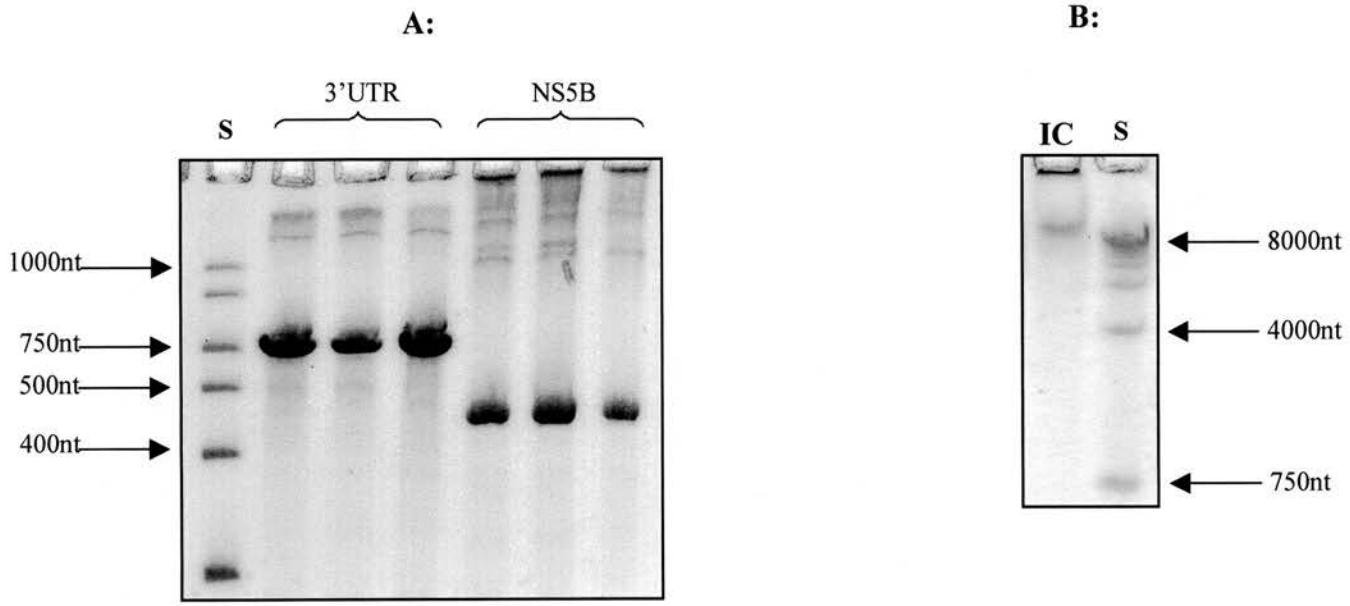
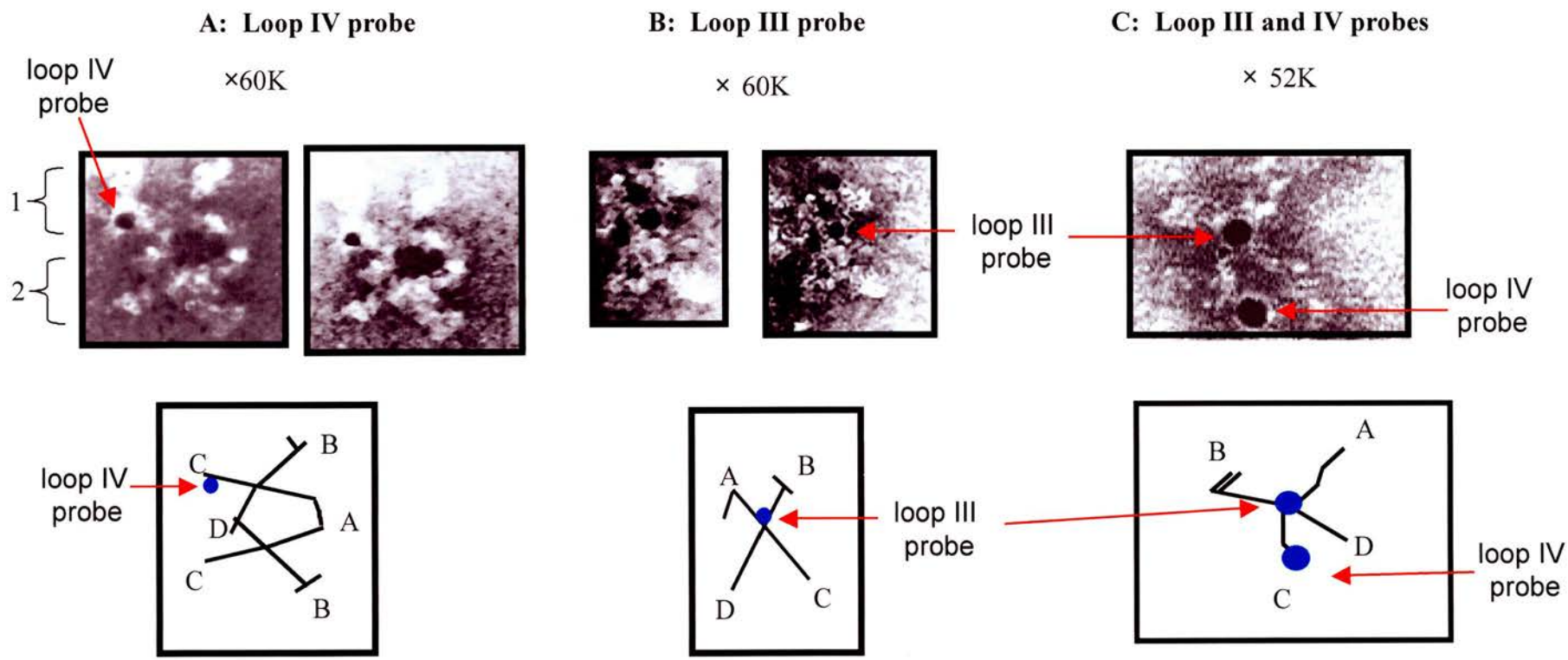


Figure 5.2. Gold labelled HGV/GBV-C NS5B RNA transcript electron micrograph images. Biotinylated oligonucleotide hybridised to single stranded domains within predicted RNA structures. **A**: Terminal single stranded loop of $SL_{NS5B}IV$ gold labelled (two micrographs of same RNA molecule showing differing qualities of resolution); two molecules visualised labelled 1 and 2. **B**: Single stranded bulge at base of $SL_{NS5B}III$ gold labelled (two micrographs of same RNA molecule taken at different resolutions). **C**: Both $SL_{NS5B}III$ and $SL_{NS5B}IV$ labelled.



B, C and D, based on similarities in confirmation, size and position within four independent micrograph images. Although the overall confirmation of the observed structures was consistent between each micrograph there was variation in the length of each stem. Based on micrograph images A to C (Micrograph A includes two structures, both of which were separately analysed) the approximate mean length of stem A was 41.7 nm (range 38-46 nm); stem B 44.7 nm (range 41-47 nm), stem C 35.3 nm (range 30-40 nm) and stem D 35.8 nm (range 34-35 nm) (Fig. 5.3). A short side structure was consistently observed branching approximately 7 nm from the distal end of stems A and B. The angles of rotation between stems A and B ranged from 57° to 67° (mean 61.7°); B and C from 122° to 123° (mean 119°); C and D 59° to 78° (mean 68°) and D and A 102° to 113° (mean 111°) (Fig. 5.3).

Gold labelling via biotinylated oligonucleotides (10 nm diameter) complimentary to either the terminal loop of RNA structure $SL_{NS5B}IV$ (loop IV probe) or the single stranded bulge towards the base of RNA structure $SL_{NS5B}III$ (loop III probe), were used to confirm that the observed structures were indeed HGV/GBV-C NS5B encoding region transcripts (complete predicted structural map shown in Fig. 4.3). Gold labelling also enabled the identification of individual stems and an approximation of size to be made. The streptavidin coated colloidal gold appeared as a dense circular mass attached to individual stem structures. The loop IV gold probe was observed hybridized towards the distal end of stem C in structure 1 of micrograph A (Fig. 5.2A). The loop III gold probe hybridised towards the proximal end of stem B, also overlaying slightly with stem A in (Fig. 5.2B). In micrograph C both the loop III and IV probes were used for RNA hybridisation (Fig. 5.2C). In this

Figure 5.3. Schematic representations of HGV/GBV-C NS5B transcript electron micrograph images, showing the angles of rotation and lengths of each of each of the visualised stems.

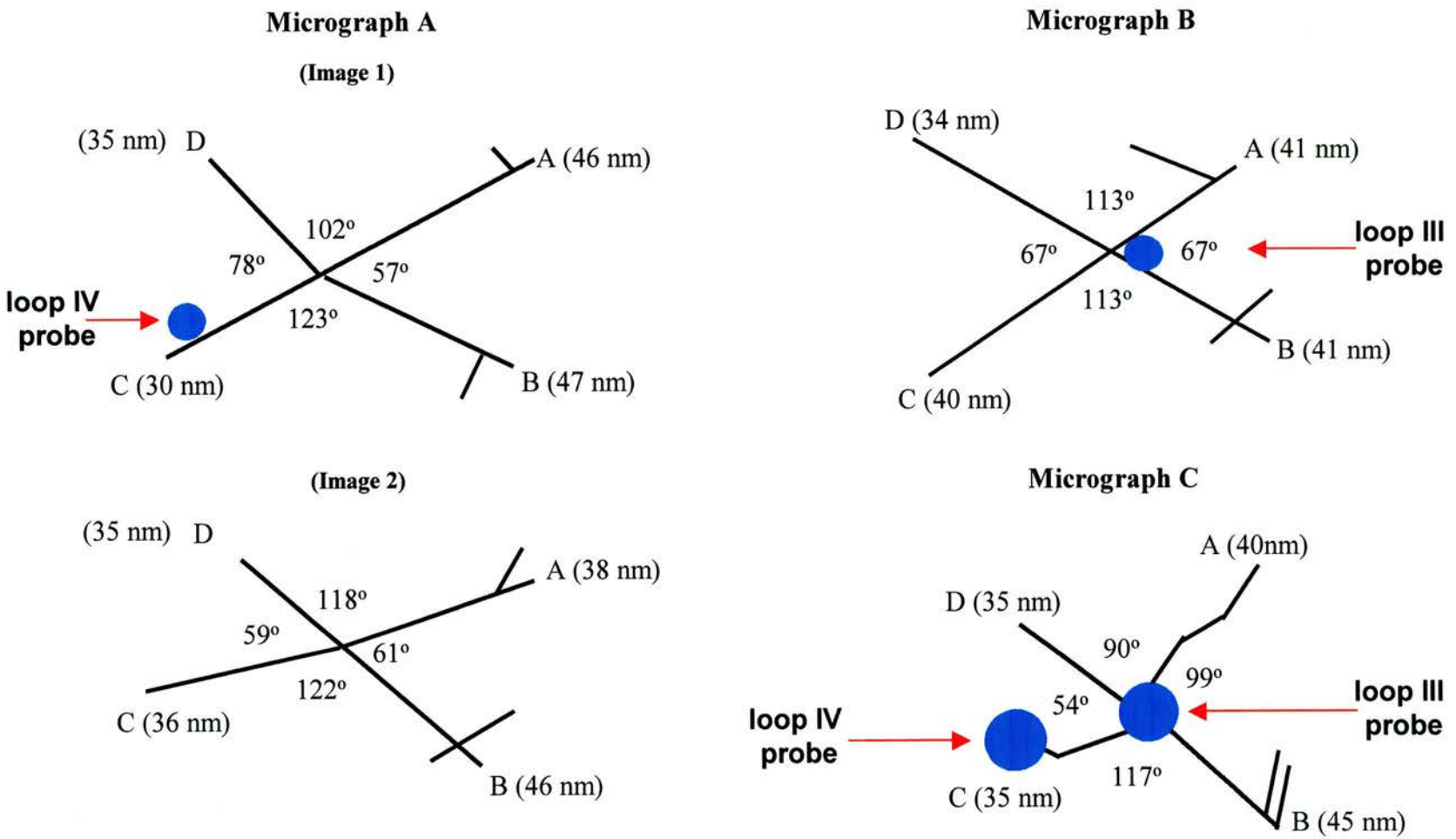


image streptavidin coated colloidal gold was observed hybridised towards the distal end of stem C and the central core of the structure.

The overall secondary structure of the downstream region of NS5B encoding RNA (8410-8894) that was visualised was also computationally predicted, using the MFOLD free energy minimisation algorithm (Fig. 5.4). This aided in the identification of individual stems visualised in the micrograph images and helped to assess the accuracy of the method. The MFOLD prediction showed a high degree of structural agreement to the micrograph images, with four stems radiating from a central axis. As with the micrograph images stems A and B were the longest structures and were located at opposing ends of the central axis to stems C and D. Stem A was proportionally longer and stem B shorter in the MFOLD prediction then the micrograph images. The structural prediction created by MFOLD is limited, as the algorithm does not take into account tertiary structure and presents a schematic diagram which does not offer a correct representation of the angles of rotation between and within predicted structures. For example, stem A was visualised as having a kink at its distal end in the micrograph images yet the computer image shows a linear stem.

Visualisation of the HGV/GBV-C 3'UTR RNA proved more problematic than for the NS5B coding region. Although, a consistently linear shaped structure was observed under the electron microscope it proved extremely difficult to resolve this structure on a micrograph image. A single resolved micrograph image, consistent with those observed by eye under the electron microscope, revealed a long structure approximately 57.7 nm in length (Fig. 5.5). The two termini, either end of the structure, were arbitrarily labelled A and B. Four short structures of approximately

Figure 5.4. Secondary structure prediction of the HGV/GBV-C NS5B region corresponding to the region visualised under the transmission electron microscope. Prediction was made using the MFOLD 3.1 free energy minimisation algorithm. Conserved RNA stem loops are labelled in brackets; biotinylated oligonucleotide hybridisation sites and labels according to electron micrograph images are shown in bold type.

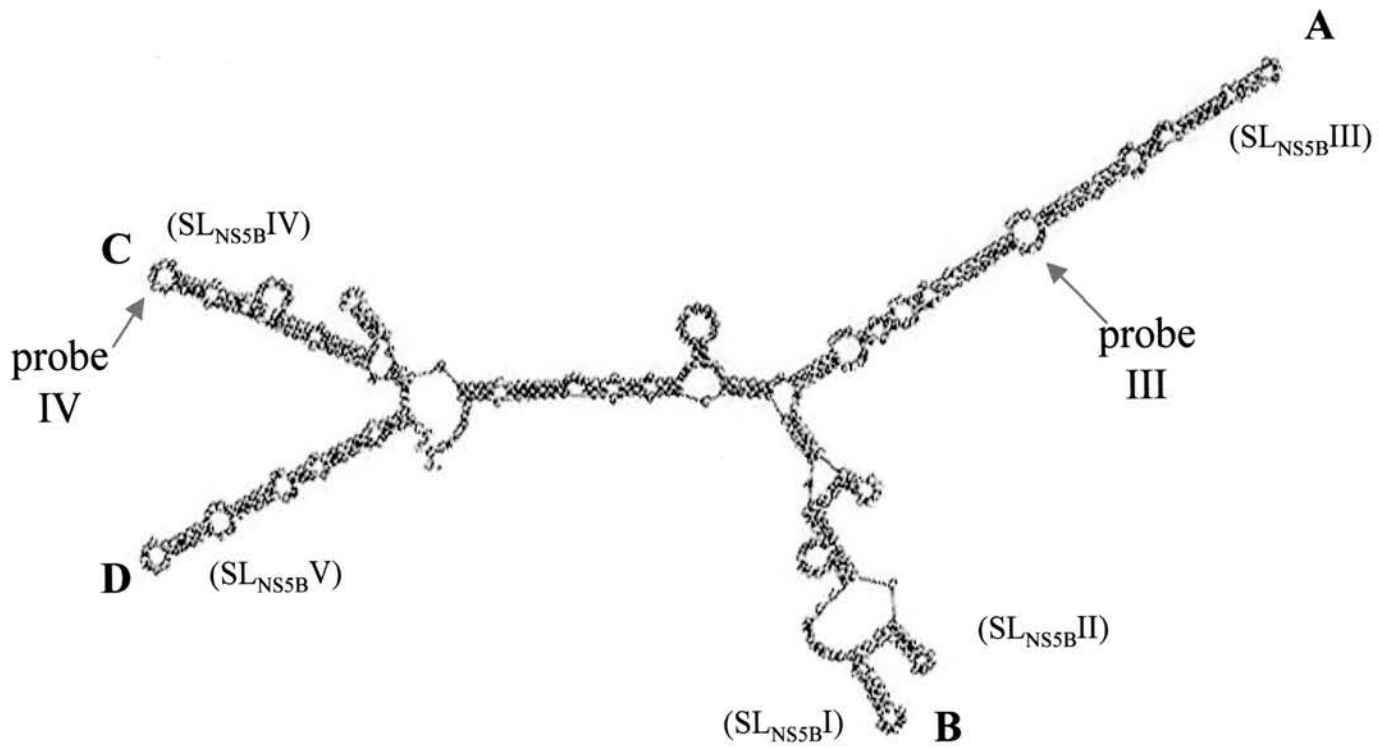
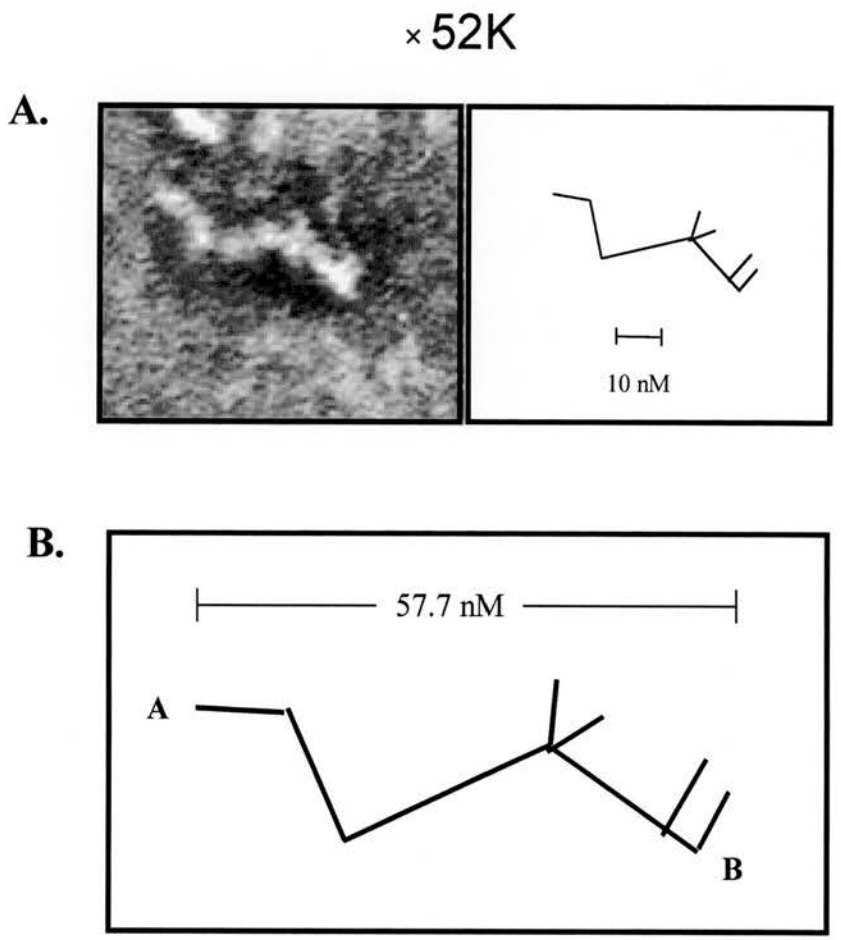


Figure 5.5. **A**: Electron micrograph image of HGV/GBV-C 3'UTR RNA transcript. **B**: schematic representation showing the approximate size of the structure.

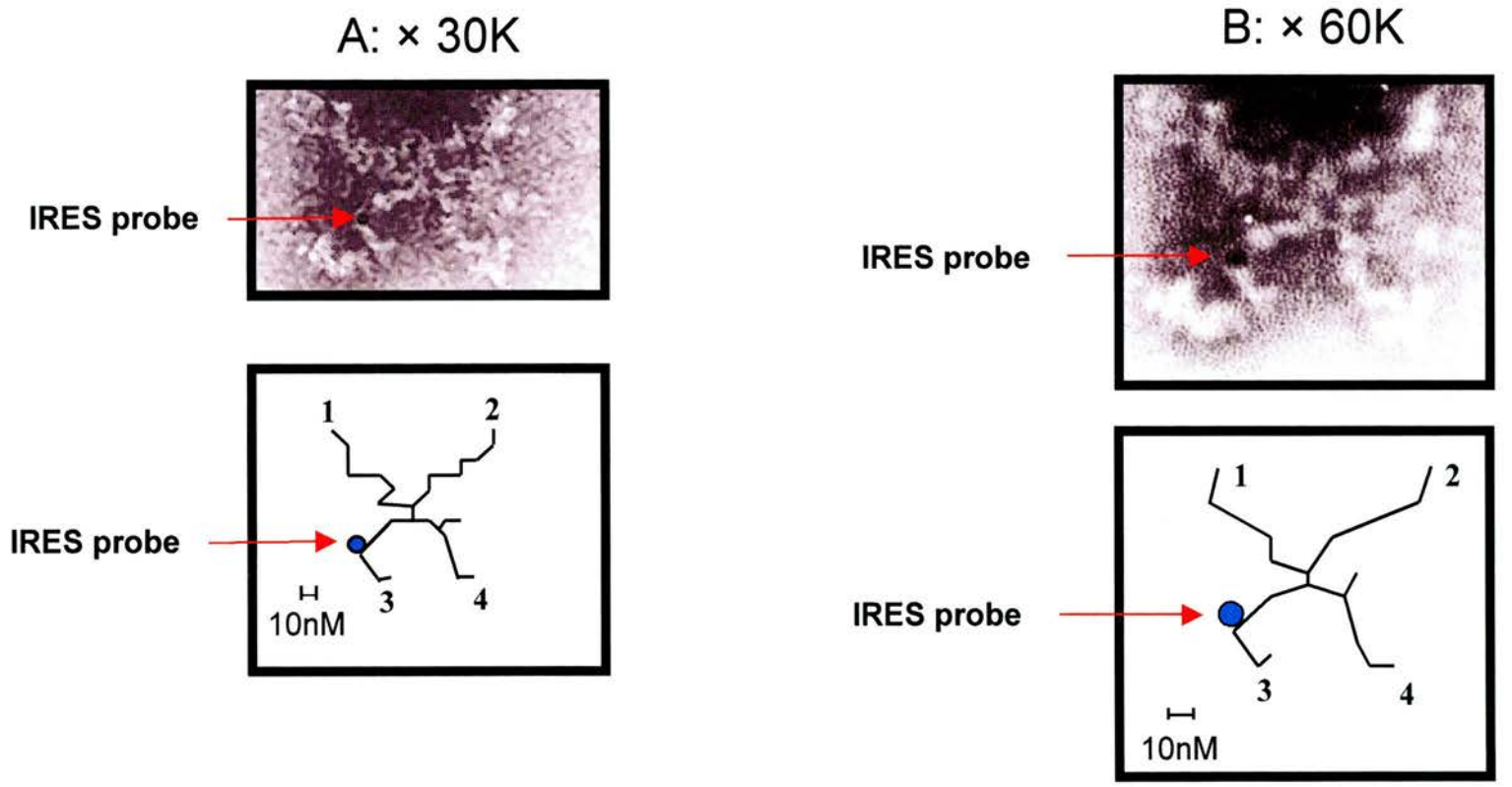


6.5 nm were observed branching from the main stem; two at terminus B and two 19 nm away from the B terminus at a 135° kink in the main stem. A further large kink in the main stem, with an angle of rotation of 76°, was noted 20 nm from terminus A. However, gold labelling via biotinylated oligonucleotides complimentary to the terminal single stranded loop of RNA structure 3'SLII failed to hybridise to this structure. Consequently, due to the lack of comparable images and the failure of colloidal gold hybridisation it was not possible to make any predictions as to the RNA folding structure of the 3'UTR region based on this study.

5.2.2 VISUALISATION OF THE COMPLETE VIRUS GENOME RNA FOLDING CONFORMATION

Examination of the transcribed HGV/GBV-C complete genomic RNA revealed a molecule consisting of four main stems radiating from a central, more clumped region (Fig. 5.6). The four stems were arbitrarily labelled 1, 2, 3 and 4 and measured approximately 67.5 nm, 72.5 nm, 57.5 nm and 66.25 nm respectively. A 10 nm long side structure was also noted branching from stem 4; 15 nm from the central core of the structure. Gold labelling via a biotinylated oligonucleotide, complementary to the single stranded bulge in domain IVa of the IRES, was observed hybridised to stem 3 and confirmed the identity of the object visualised in the electron micrographs. The colloidal gold label was observed bound half way up stem 3, in the region of a 101.5° kink in the structure.

Figure 5.6 Gold labelled HGV/GBV-C complete genome transcript electron micrograph images. Domain IVa of the IRES within the 5'UTR labelled with streptavidin-coated colloidal gold. Individual stems arbitrarily labelled 1, 2, 3 and 4 within schematic representations. Both micrographs are of the same RNA molecule taken at different magnifications. **A**: Micrograph taken at ×30 000 magnification. **B**: Micrograph taken at ×60 000 magnification.



Two electron micrographs of the same molecule were resolved, one at ×30 000 magnification (A) and a second at ×60 000 magnification (B). The resolution of the lower magnification image was the better of the two, revealing a fifth possible structure between stems 2 and 4. However, it was not possible to accurately visualise this potential stem at the lower magnification used in micrograph A. In micrograph B the region was obscured due to a loss of resolution which was observed at higher magnification. The central region of the structure appeared to be composed of a number of discrete strands, although these were not fully resolved in either electron micrograph and so it was not possible to accurately map this region.

5.3 DISCUSSION

Previously, a combination of thermodynamic and phylogenetic methods were used to predict and map stable RNA stem loop structures within both the NS5B encoding region and 3'UTR of the HGV/GBV-C genome (chapters 3 and 4) (Cuceanu et al., 2001; Simmonds and Smith, 1999b; Katayama et al., 1998; Okamoto et al., 1997). We also provided evidence suggesting that the HGV/GBV-C may be more extensively folded, across the complete length of its genome (chapter 3) (Cuceanu et al., 2001). In this study negative staining followed by transmission electron microscopy and gold label hybridisation has been used to directly visualise and identify five RNA structures within the NS5B region (Fig. 5.2). These structures were consistent with the downstream stem loops previously mapped within the NS5B encoding region of HGV/GBV-C ($SL_{NS5B}V$ to $SL_{NS5B}I$). 3'UTR RNA transcripts were also visualised, although it proved difficult to resolve these micrograph images (Fig. 5.5). Visualisation of the complete virus genome transcript revealed a highly structured molecule consistent with local and long range RNA structure along the length of the HGV/GBV-C genome (Fig. 5.6).

Methods of RNA structure visualisation such as X-ray crystallography (Robertus et al., 1974), cryoelectron microscopy (Gabashvili et al., 1999), and nuclear magnetic resonance (NMR) (Lukavsky et al., 2000; Montelione et al., 2000), have previously been used to provide very detailed, atomic level, RNA structure information. However, all of these methods require relatively short RNA fragments or ribonuclear protein complexes to aid crystal formation. Both the NS5B region and the complete

genome RNA transcripts of HGV/GBV-C are relatively long, naked RNA molecules. Consequently, it was decided that negative staining followed by transmission electron microscopy would be used to directly visualise the structure of HGV/GBV-C RNA transcripts, as sample length is not prohibitive with this method. Negatively stained electron micrograph images of RNA virus IRES structures have previously been published (Beales et al., 2001; Beales et al., 2003; Yunoki et al., 2003), although there are no examples in the literature of this method having been used to investigate the structure of virus coding region RNA or complete virus genomes.

Formerly direct visualisation of RNA structure by electron microscopy has relied upon coating RNA with spermidine, or other polyamines, followed by rotary shadow-casting with heavy metals such as platinum (Wang et al., 1994; Nakamura et al., 1995). Spermidine has since been shown to alter the structural conformation of a ribozyme four way junction (Walter et al., 1998), making it impractical for the analysis of RNA folding structure. Negative staining is not believed to affect RNA conformation in this way and also results in a smaller grain size than heavy metal shadowing, increasing the resolution of the micrograph images (Beales et al., 2001).

5.3.1 VISUALISATION OF HGV/GBV-C NS5B REGION RNA FOLDING STRUCTURE

The negatively stained NS5B region was observed to fold into a cruciform conformation (Fig. 5.2), which was consistent with a computational prediction of overall confirmation created using free energy minimisation (Fig. 5.4). Based on

conformational similarities to the predicted structure and gold labelling of specific stem loops it is hypothesised that stem A corresponds to $SL_{NS5B}III$, B to stems $SL_{NS5B}I$ and $SL_{NS5B}II$, C to $SL_{NS5B}IV$ and D to $SL_{NS5B}V$ (complete predicted structural map shown in Fig. 4.3). It is possible to make an estimate of the length of a predicted stem loop based on the assumption that RNA nucleotide pair stacks adopt an A form helix in which paired nucleotides rise at 0.273 nm per base and lateral side chains at 2.5 nm (Dock-Bregeon et al., 1989; Beales et al., 2001). Following these parameters the computationally mapped stem A structure was estimated to be 16.7 nm in length, B 11.82 nm, C 12.64 nm and D 9.56 nm. This compares to an average length determined from the micrograph images of 41.7 nm, 44.7 nm, 35.3 nm and 34 nm respectively (Fig. 5.3). The underestimation in predicted stem loop length, compared to the directly visualised structures, may be due to the presence of single stranded bulges which are present in each of the predicted structures and are of an indeterminate length (Beales et al., 2001).

There is close agreement in the individual stem loop angles of rotation between separate micrograph images (Fig. 5.3). The angles between stems C to B (mean 119°) and A to D (mean 111°) were consistently greater than those between C to D (mean 68°) and A to B (mean 61.7°). However, in micrograph C the angle of rotation between stems A to B was larger (99°) than that between A to D (54°). The structure in micrograph B was labelled at two separate points with colloidal gold which may have disrupted individual angles of rotation. It has previously been proposed that stem loop flexibility is required for pseudoknot formation in the IRES of HCV (Nakamura et al., 1995; Beales et al., 2001). The variation in angles observed within the NS5B encoding region of HGV/GBV-C may thus suggest a level of flexibility in

stem loop conformation in order to allow tertiary structure formation; in which stem flexibility would enable the interaction of unpaired nucleotides.

Hybridisation with biotinylated oligonucleotides followed by labelling with streptavidin coated colloidal gold confirmed that the negatively stained image was that of the NS5B encoding region RNA transcript (Fig. 5.2). Gold labelling also confirmed the identity of stems A and C as RNA structures $SL_{NS5B}III$ and $SL_{NS5B}IV$ respectively. Comparison with a free energy minimisation predicted structure for the NS5B encoding region also allowed the identity of stem D to be surmised as RNA structure $SL_{NS5B}V$ and the two branched structures at the distal end of stem D as RNA structures $SL_{NS5B}I$ and $SL_{NS5B}II$. However, without specific probes for these stem loops it was not possible to make a more definitive prediction or show which of the distal branches was RNA structure $SL_{NS5B}I$ or $SL_{NS5B}II$.

Phylogenetic and thermodynamic methods previously predicted stem loop $SL_{NS5B}IV$ to be only twelve nucleotide pairs in length as compared to stem loops $SL_{NS5B}V$ and $SL_{NS5B}III$ which are respectively thirty eight and thirty four nucleotides in length. However, both $SL_{NS5B}III$ and $SL_{NS5B}IV$ are observed as being longer than $SL_{NS5B}V$ in the electron micrographs images. This suggest that although, the phylogeneticaly conserved regions of the two stem loops are shorter than $SL_{NS5B}V$ they are presented on a long but less sequence dependent proximal base paired stalk.

5.3.2 VISUALISATION OF THE HGV/GBV-C 3'UTR RNA FOLDING STRUCTURE

Resolution of a HGV/GBV-C 3'UTR electron micrograph image of RNA transcripts proved difficult. This may have been due to the fact that less RNA structure was predicted within the 3'UTR than the NS5B coding region, making the RNA molecules less easy to resolve against the negatively stained background. Less RNA structure may also have meant that the 3'UTR RNA transcripts were more prone to digestion with contaminating ribonucleases. Although, a single micrograph image revealed a structure approximately 58 nm in length, gold label hybridisation failed to confirm the identity of this structure. Consequently, it was not possible to make specific predictions concerning the RNA folding structure of the HGV/GBV-C 3'UTR based on direct imaging via electron microscopy.

5.3.3 VISUALISATION OF THE HGV/GBV-C COMPLETE GENOMIC RNA FOLDING STRUCTURE

Direct visualisation of the HGV/GBV-C complete genome revealed four main negatively stained stems radiating from a central core (Fig. 5.6). The central region was less well resolved than the radiating structures but appeared to be composed of discreet individual structures. Hybridisation with a biotinylated oligonucleotide, specific for domain IVa of the HGV/GBV-C IRES, followed by labelling with streptavidin-coated colloidal gold confirmed that the negatively stained image was that of the HGV/GBV-C full length RNA transcript.

The complete genome of HGV/GBV-C (9231 nucleotides) is approximately twenty times the length of the NS5B encoding region visualised in this study (484 nucleotides). However, the diameter of the complete genome structure is only approximately twice that of the NS5B encoding region. The diameter of the observed complete genome structure was 133.75 nm across the axis of stems 1 to 4 and 138.75 nm across the axis of stems 2 to 3 (Fig. 5.6). The average diameter of the NS5B region RNA structure was 76.7 nm across the axis of stems A to C and 79.2 nm across the axis of stems B to D (Fig. 5.3). This apparent contradiction suggests that RNA structure may not be confined to the extreme ends of the genome, or indeed any to any discreet region, but that the complete genome may be folded through local or long range interactions to form a compact, highly structured, tertiary conformation. This is consistent with the thermodynamic and phylogenetic results presented in chapter 3 in which extensive RNA structure was predicted to exist across the polyprotein coding region. Similar results have been noted for the complete genome of HCV, in which three main stems are observed radiating from a central core region (Lucy Beales, personnel communication).

5.3.4 ROLE OF SECONDARY STRUCTURE IN HGV/GBV-C

The functional role of RNA structure within the NS5B coding region and 3'UTR of HGV/GBV-C has yet to be identified. In similar positive sense, single stranded, RNA viruses a number of stem loops have been identified within the polyprotein coding region which function as *cis*-acting replication elements (CRE). These

structures have been shown to be essential for negative strand initiation in a number of viruses including enteroviruses (Goodfellow et al., 2000), cardioviruses (Lobert et al., 1999), aphthoviruses (Mason et al., 2002) and rhinoviruses (McKnight and Lemon, 1998) (described in more detail in section 6.3). It may be that the RNA structures identified in this study within the NS5B encoding region of HGV/GBV-C perform a similar role in virus replication, although they are larger and more extensive than any previously predicted in similar viruses.

RNA structures identified within the 3'UTR may have an analogous role to similarly structured regions predicted in related virus, which play a critical role in the initiation of both positive and negative strand replication, through specific RNA-protein interactions. For example, stem loop structures within the 3'UTR of West Nile virus (WNV) act as specific recognition sites for host cellular proteins such as EF-1α (Blight and Rice, 1997; Blackwell and Brinton, 1995; Shi et al., 1996). The 3'UTR of HCV has also been shown to be absolutely required for virus replication and has been shown to bind a number of replication complex proteins such as PTB RdRp (NS5B protein) (more detail in section 6.3) (Yanagi et al., 1999; Banerjee and Dasgupta, 2001; Ito and Lai, 1997).

The close similarities in excess free energy on folding the complete genomes of plant viroids, the non-coding region of delta viruses and HGV/GBV-C have previously been noted (chapter 3) (Cuceanu et al., 2001). In viroid agents and the non-coding region of delta virus RNA folding is essential for replication of the genome in which specific domains catalyse RNA cleavage, editing and sequence ligation (Branch and Robertson, 1984). The complete genome folding of HGV/GBV-C may play a similar role or be involved in genome packaging or

protection against ribonuclease enzymes particularly, since HGV/GBV-C does not appear to encode a nucleocapsid protein.

In summary, negative staining followed by transmission electron microscopy has been used to visualise the folding structure of the NS5B region and complete genome of HGV/GBV-C. The visualised structures are consistent with phylogenetic and thermodynamic results presented earlier in chapters 3 and 4. Whilst the function of the RNA structures predicted within the HGV/GBV-C genome are currently not understood in any detail these results may provide a starting point for further functional studies using the HGV/GBV-C replicating full length transcript system (Xiang et al., 2000).

CHAPTER 6

RIBONUCLEASE MAPPING OF RNA STRUCTURE

6.1 INTRODUCTION

Ten evolutionary conserved stem loop structures were previously predicted within the coding region of HCV, using both thermodynamic and phylogenetic methods (chapters 3 and 4) (Tuplin et al. 2002). Three were predicted and mapped within the core gene and seven within the NS5B region. Each structure was shown to be highly conserved between divergent genotypes with multiple covariant and semi-covariant substitutions observed in each structure, suggesting functional conservation (chapter 4).

The free energy minimisation methods used in chapter 4 do not take account of tertiary structures, such as long range interactions, which may have a functional role and affect stem loop stability and structure. For example, evidence from HCV IRES driven translation in a dicistronic reporter construct, suggests that the core gene may play a role in the modulation of virus translation (Reynolds et al 1995; Kim et al, 2003; Honda et al., 1999). It was shown that this function may be dependent upon a long range interaction between the core gene and the 5'UTR. A potential core gene long range interaction site was variously mapped between nucleotides 87 to 101 (Kim et al., 2003), and/or 66-588 (Honda et al, 1999). The three core gene RNA structures predicted in chapter 4 (SL47, SL87 and SL443) fall within this region and may thus play a role in higher order interactions affecting the regulation of translation. However, although such tertiary interactions may affect RNA structure they are not accounted for in free energy minimisation predictions.

In order to examine the stem loops predicted in HCV in as natural a context as possible the core gene and NS5B region were enzymatically mapped using a multisite priming method which allowed the analysis of long RNA transcripts. The use of long transcripts enabled potential tertiary interactions between secondary structures to occur prior to cleavage. Both genotypes 1 and 2 were analysed in order to assess the conservation of ribonuclease cleavage patterns between divergent sequences

SEQUENCES ANALYSED

The RNA transcripts were transcribed *in vitro* using T7 phage promoter annealed PCR product as a template. The core region templates corresponded to positions -23 to 516 and the NS5B regions 8180 to 8715 of full length, functional HCV infectious clones of genotype 1 (M62321) and 2a (AF177036).

6.2 RESULTS

Through analysis of variability at synonymous sites, analysis of covariant substitutions, thermodynamic prediction and calculation of excess free energy on folding we have previously obtained evidence for extensive RNA secondary structure within the coding region of HCV (chapters 3 and 4) (Smith and Simmonds, 1997a; Tuplin et al., 2002). Three stem loops were predicted within the core gene and seven within the NS5B region (Figs 4.1 and 4.2).

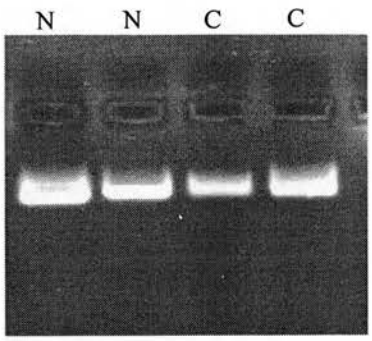
In this study a multisite RNA transcript priming method was developed to enzymatically map the complex and extensive predicted RNA structures in the core and NS5B regions of the HCV (section 2.5). Long RNA transcripts (approximately 500 bases) were used as templates for ribonuclease digestion as it was essential to allow stem loop formation to occur in as natural a context as possible and in particular to enable potential interactions between RNA secondary structures, such as formation of pseudoknots, to occur prior to nuclease cleavage.

RNA was generated by *in vitro* transcription using a MegaScript kit (Ambion) and T7 phage polymerase promoter annealed PCR product as a template (section 2.3). The PCR product templates corresponded to positions -23 to 516 within the core gene and 8180 to 8715 within the NS5B encoding region of full length, functional HCV infectious clones of genotypes 1a and 2a (section 2.1.3). The integrity of the RNA was determined by electrophoresis through a 1% agarose gel (typical example is shown in Fig. 6.1). Prior to cleavage the RNA was melted and slowly annealed under physiological ionic conditions. RNase partial digestion was

performed with ribonucleases V1, T1 or A1, which preferentially cleave downstream of base-pared regions, unpaired G or unpaired C and U nucleotides respectively (section 2.5).

Three antisense primers were used for transcription (in the presence of an α-[^{33}P]ATP substrate) from multiple sites in the transcripts to analyse consecutive regions of the RNA cleavage products of the core gene (corresponding to positions 157-176, 327-456 and 497-516) and NS5B encoding region (corresponding to positions 8961-8981; 9061-9080 and 9161-9180). The radio-labelled cDNA fragments were separated by electrophoresis through a 5% denaturing acrylamide gel, dried and exposed on autoradiography film (typical examples are shown in Fig. 6.2).

Figure 6.1. Determination of RNA transcript integrity, for HCV core gene (C) and NS5B region (N), after electrophoresis through a 1% agarose gel. RNA is stained with 0.07 μg/ml ethidium bromide (section 2.3.5). RNA transcripts from 2 independent transcription reactions are shown for each region

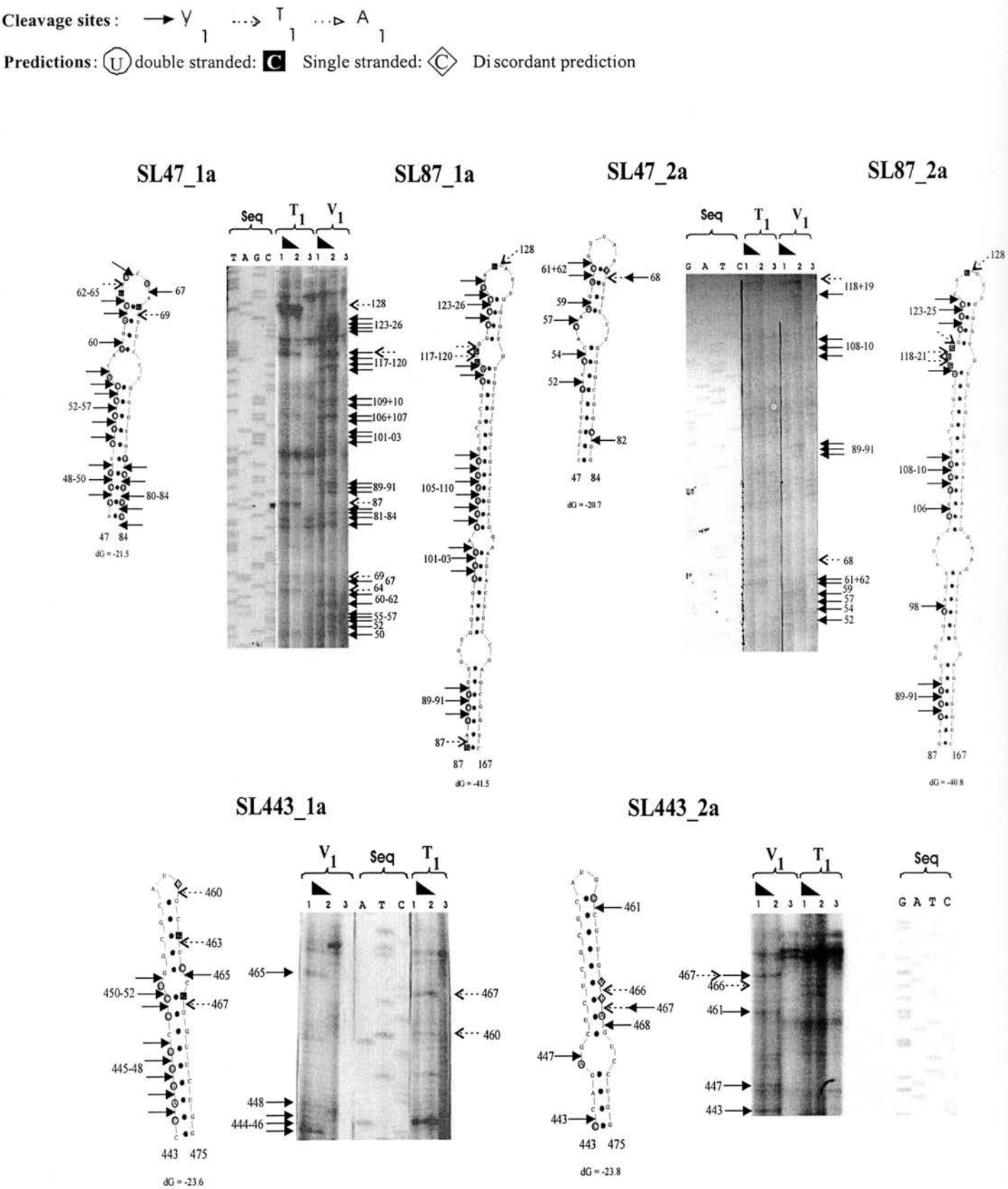


6.2.1 HCV CORE GENE ENZYMATIC MAPPING

Three evolutionarily conserved stem loops, SL47, SL87 and SL443, were previously predicted in the core gene region of HCV, spanning regions 47 to 84; 87 to 167 and 443 to 475 respectively (Fig. 4.1) (chapter 4) (Tuplin et al., 2002). Both the position and structural conservation of these stem loops were consistent with the cleavage patterns of the RNA transcripts in the present ribonuclease mapping study (Fig 6.3). As well as locating the stem loop structures, the ribonuclease cleavage results were consistent with the predicted lengths and positions of the base paired stems, the location of the terminal single stranded regions and many of the single stranded bulges found within the main paired stem structures. Furthermore, the cleavage patterns obtained by RNAse mapping were very similar between genotype 1a and 2a transcripts, despite the often high degree of primary sequence divergence between the genotypes.

While the nuclease mapping data generally corresponded closely to the optimally folded structures from MFOLD for genotypes 1a and 2a, there were some instances where the two methods made discordant predictions (Fig. 6.3). For example, the terminal loop of SL47 was previously predicted to be single stranded and the results from ribonuclease T_1 confirm this. However, the results from V_1 suggest that two out of the five bases in the terminal loop of SL47_1a are paired (positions 65 and 67). Within the same stem loop a single stranded bulge is predicted by computational analysis, between positions 57 to 59 and 73 to 74, V_1 enzymatic cleavage indicated that positions 57 and 58, within genotypes 1a and 2a respectively, are paired. Positions 452 and 460 within SL443 (genotype 1a) also highlight

Figure 6.3. Representative ribonuclease cleavage results alongside schematic representations of stem loops indicating enzymatic cleavage positions within the core gene of HCV (genotypes 1a and 2a indicated by bold type. Cleavage sites by the nucleases V_1, A_1 and T_1 are indicated by solid, hollow and outline arrows respectively (see Figure key). Each RNA transcript was cleaved with two different concentrations of each RNAse, and a no nuclease control to detect non-specific cleavage or transcription pausing during strand extension (blocks 1, 2 and 3 respectively). Fragments were sized by comparison wi1th a cycle sequencing reaction of a DNA template of the same sequence (block Seq).



instances in which the cleavage results contradict the computational predictions; both nucleotides lie within predicted single stranded regions yet V_1 cleavage results suggest that they are paired. Interestingly position 460 also maps as single stranded with ribonuclease T_1. It has been suggested that this region of the core gene modulates translation through the formation of local higher order structures and a transient long range pseudoknot structure with the upstream IRES (Kim et al., 2003; Rijnbrand et al., 2001; Honda et al., 1999). Such interactions may account for these apparent contradictions between the computational and ribonuclease cleavage results, as most computational methods are unable to predict higher order structures such as pseudoknots.

The results also indicate that paired nucleotides adjacent to a single stranded region may map enzymatically as both single stranded and paired. This is the case for nucleotide 69 (in both genotypes 1a and 2a) which is directly adjacent to the terminal single stranded loop in SL47 and nucleotide 67 within SL443 (genotype 1a). However, these results are consistent with the known dynamic instability of stacked base pairs adjacent to unpaired regions (Zuker, 2000).

6.2.2 HCV NS5B REGION ENZYMATIC MAPPING

Ribonuclease cleavage patterns were also consistent with the structure and position of stem loops SL8828; SL8926; SL9011; SL9061 and SL9118 spanning positions 8828 to 8897; 8926 to 8987; 9011 to 9059; 9061 to 9106 and 9118 to 9148 respectively within the NS5B protein encoding region (Fig. 6.4). All five stem loops

Figure 6.4. Representative ribonuclease cleavage results alongside schematic representations of stem loops indicating enzymatic cleavage positions within the NS5B region of HCV (genotypes 1a and 2a indicated by bold type. Cleavage sites by the nucleases V_1, A_1 and T_1 are indicated by solid, hollow and outline arrows respectively (see Figure key). Each RNA transcript was cleaved with two different concentrations of each RNAse, and a no nuclease control to detect non-specific cleavage or transcription pausing during strand extension (blocks 1, 2 and 3 respectively). Fragments were sized by comparison wi1th a cycle sequencing reaction of a DNA template of the same sequence (block Seq).

Cleavage sites : → V ⇢ T ▷ A

Predictions : (U) double stranded: **C** Single stranded: ◇C Di scordant prediction

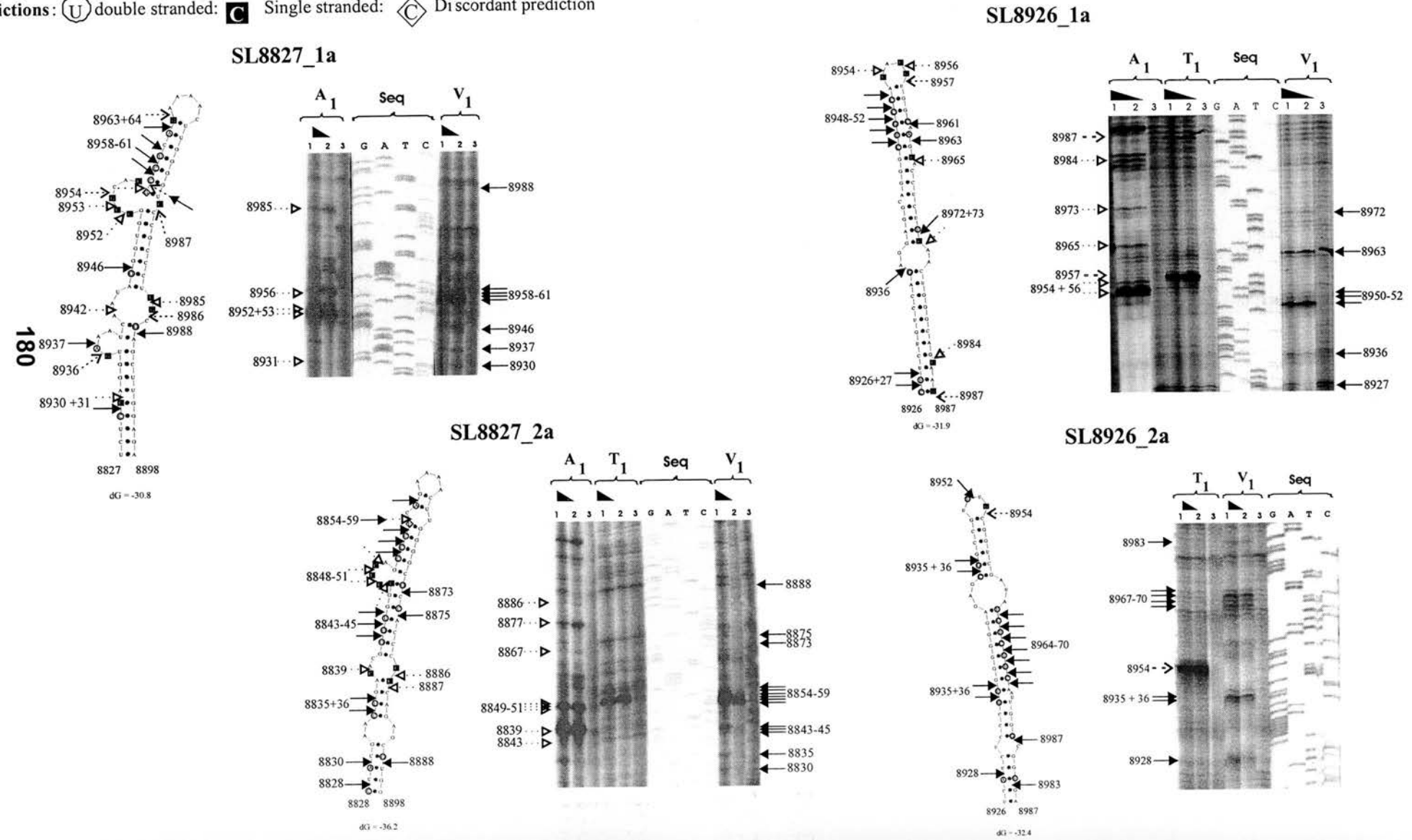


Figure 6.4. Representative ribonuclease cleavage results alongside schematic representations of stem loops indicating enzymatic cleavage positions within the NS5B region of HCV (genotypes 1a and 2a indicated by bold type. Cleavage sites by the nucleases V_1, A_1 and T_1 are indicated by solid, hollow and outline arrows respectively (see Figure key). Each RNA transcript was cleaved with two different concentrations of each RNAse, and a no nuclease control to detect non-specific cleavage or transcription pausing during strand extension (blocks 1, 2 and 3 respectively). Fragments were sized by comparison wi1th a cycle sequencing reaction of a DNA template of the same sequence (block Seq).

Cleavage sites : → V ⇢ T ▷ A

Predictions : (U) double stranded: **C** Single stranded: ◇C◇ Di scordant prediction

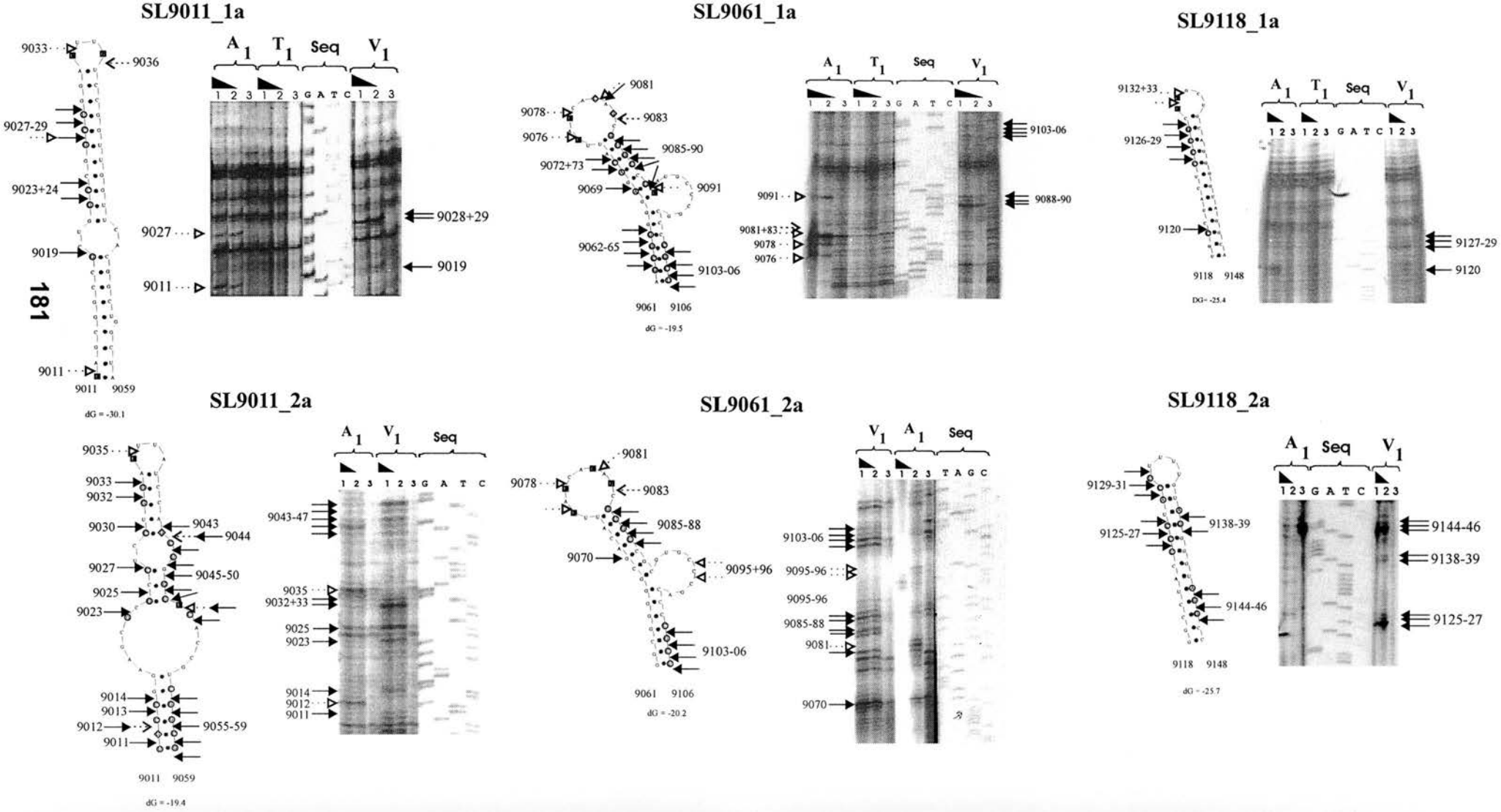


exhibit a high degree of structural conservation between genotypes 1a and 2a, following both the thermodynamic/phylogenetic methods of prediction and ribonuclease cleavage patterns.

Intriguingly, as with the core gene, there are a number of instances when the two predictive methods appear to contradict each other. For example, nucleotides 8837, 8954, 9047, 9081 and 9031 within stem loops SL8828 (genotype 1a), SL8926 (genotype 2a), SL9011 (genotype 1a), SL9061 (genotype 1a) and SL9118 (genotype 2a) respectively were observed in stranded domains with both computational and ribonuclease (T_1 and/or A_1) cleavage results. However, all five positions also mapped as paired in ribonuclease V_1 cleavage results. This may be due to some of the apparently unpaired bulges within the stem loops forming pseudoknot like interactions with other structures or with completely unpaired regions of the virus genome. Evidence for this, may be the high degree of sequence conservation in unpaired regions, which was observed even at synonymous sites such as in the terminal stem loop of SL9061.

Paired bases 8858 –8876 and 8864 – 8870 within SL8828 (genotype 1a); 8842 – 8885, 8852 – 8877 and 8891 – 8872 within SL8828 (genotype 2a); 8926 – 8987, 8939 – 8973 and 8947 – 8965 within SL8926 (genotype 1a); 9012 – 9058 and 9028 – 9041 within SL9011 (genotype 1a); 9011 – 9059 within SL9011 (genotype 1a) and 9068 – 9091 within SL9061 (genotypes 1a and 2a); were all part of stacked base pairs adjacent to single stranded regions which mapped by ribonuclease cleavage as unpaired or as both paired and unpaired. This is in marked contrast to contradictory results towards the middle of base pair stacks, which was only observed once within the NS5B region, at paired nucleotides 8831 – 8894 in SL8828 (genotype 1a). These

results provided further evidence for increased instability of stacked base pairs adjacent to single stranded regions. Instances when the results from the various methods appear to contradict each other within base paired stacks may be evidence that the RNA stem loop, or a particular domain within it, is inherently unstable and able to adopt multiple conformations (Goodfellow et al 2003).

6.3 DISCUSSION

Although extensive stable stem loop structures have been predicted within the 5' and 3' untranslated regions (5' and 3' UTRs) of single stranded RNA viruses, very little is known about possible structures within the coding region of HCV and related viruses. In this study, ribonuclease cleavage followed by multisite transcript priming was used to investigate the presence and elucidate the structure of a large number of stable stem loops within the coding region of HCV. Physical mapping also enabled the validation of both thermodynamic and phylogenetic predictive methods which were previously used (chapter 4) (Tuplin et al. 2002; Smith and Simmonds 1997b).

In chapter 4 a variety of both thermodynamic and phylogenetic methods were used to predict ten evolutionarily conserved stable stem loops within the positive sense RNA genome of HCV. Ribonuclease cleavage mapping of regions of predicted structure within HCV formed an important part of this study as it allowed the validation of these computational methods. Enzymatic mapping also supplied additional information which was not highlighted by computational predictions. For example, a number of single stranded regions within stem loops were observed to base-pair, potentially indicating the formation of long range RNA-RNA interactions or higher order structures such as pseudoknots. The dynamic nature of some stem loops, in which bases may be either paired or unpaired, was also highlighted. This was shown to be more common towards the extremes of stacked regions such as adjacent to unpaired bulges or the base of a structure.

A number of other physical mapping methods could have been employed to validate the computational methods used in chapter 4. These include nuclear

magnetic resonance spectroscopy (NMR spectroscopy) and direct visualisation by transmission electron microscopy; this was used in chapter 3 when analysing the higher order structure of HGV/GBV-C. Direct mapping by ribonuclease cleavage followed by multisite primer extension holds a number of advantages over these techniques for the purposes of this study.

The regions of RNA analysed here by enzymatic cleavage were approximately 500 nucleotides in length. The fragment length required for NMR spectroscopy (100 – 200 bases) would not have been suitable in this regard, as the shorter fragment length would not have allowed the formation of large structures or potential interactions between structures to occur. The use of shorter fragments would also have given a self determining aspect to the results. For example, if regions 1 to 100 or 8900 to 9000 were analysed it is likely that both SL47 and SL8926 would have formed, however this would not have been evidence for their existence within the longer fragments (Tuplin et al. 2002). This could also be true for the enzymatic cleavage method applied here; although the potential is minimised as a results of using much longer RNA fragments, which allow potential long range interactions between structures to occur.

Transmission electron microscopy could have been used to directly visualise the folding pattern of the complete virus genome. Unfortunately, the lower resolution of this method determines that shorter stem loops would be very difficult to visualise. The two dimensional nature of the electron micrograph image may also have meant that individual stem loops would have been obscured (chapter 5).

Translation of the HCV polyprotein is directed by an internal ribosome entry site (IRES) which is located within the 5' UTR of the virus genome and is composed of

four major structural domains (Tsukiyama Kohara et al. 1992; Wang et al. 1993; Brown et al. 1992; Smith et al. 1995). It has been suggested that the activity of the IRES is modulated by a long range RNA-RNA interaction between nucleotides 87 to 101 within the core gene and -317 to -302 within the 5'UTR (referred to as long range interaction sites) (Kim et al. 2003). Mutations within nucleotides -317 to -302, which have been observed to destabilise the long range interaction, significantly enhance IRES mediated translation within a dicistronic reporter system (Kim et al. 2003). It has also been shown that a deletion between nucleotides -309 to -304 of the 5'UTR enhances IRES driven translation but only when nucleotides 66 to 588 of the core gene are included in the construct (Honda et al. 1999). Further analysis, with frame shift mutations, has shown that the nucleotide sequence of the core gene and not the amino acid sequence of the translation product is responsible for modulating IRES efficiency (Honda et al. 1999).

Intriguingly SL47, SL87 and SL443 all fall within nucleotides 66 and 588, which have been shown to modulate IRES translational activity (Kim et al. 2003; Honda et al. 1999). The 3' terminus of SL47, at nucleotide 84, directly precedes the 5' boundary of the predicted long range interaction site within the core gene (nucleotides 87 to 101), which also overlaps with the first 15 nucleotides of SL87. The close proximity of SL47 to the long range interaction site suggests that this stem loop may play a role in the formation of a higher order structure, between the core gene and the extreme 5' end of the virus genome. This could occur through SL47 exposing or presenting the long range interaction site within SL87, facilitating an interaction between the core gene and the extreme 5' end of the virus genome. The ribonuclease cleavage results presented here provide evidence that nucleotides within

the single stranded bulge of SL47 (genotypes 1a and 2a) and the single stranded terminal loop of genotype 1a also form paired structures, which is consistent with SL47 taking part in further tertiary interactions.

It has been proposed that differences in RNA structure between virus genotypes may confer differing phenotypes (Honda et al. 1999). The 5'UTR of HCV genotype 1a directs cap-independent translation with approximately twice the efficiency of genotype 1b (Honda et al. 1996). It has been suggested that this is due to a 9 nucleotide sequence difference within the 5'UTR which stabilises a long range RNA-RNA interaction (in genotype 1b) between the 5'UTR and nucleotides 67 to 588 of the core gene, down regulating IRES function (Honda et al. 1999). Ribonuclease cleavage analysis highlights a number of differences between the stem loops of genotypes 1a and 2a. For example, the single stranded terminal loops of structures SL47 and SL443 in genotype 1a show evidence of transient base pairing, which is not the case for genotype 2a. It is possible that such differences may be evidence of differing long range interactions which could affect virus phenotype.

An important aspect of the HCV replication cycle is that RNA transcription is confined to the virus genome and does not results in the amplification of cellular messenger RNAs (McKnight and Lemon 1998). For many positive-sense single stranded RNA viruses, such as rhinovirus type 14, specificity is believed to be a result of unique interactions between the replication complex proteins, RNA structures within the 5' and 3' UTRs and *cis*-acting replicating elements (CRE) within the coding sequence of the virus (McKnight and Lemon 1998). A number of *cis*-acting elements have been described in positive-sense RNA viruses including enteroviruses (Goodfellow et al. 2000), cardioviruses (Lobert et al. 1999),

aphthoviruses (Mason et al. 2002) as well as rhinoviruses (McKnight and Lemon 1998).

Electrophoretic mobility shift and competition assays have revealed that recombinant NS5B protein (RdRp) preferentially binds the 3' end of the NS5B protein coding region, between positions 8916 and 9115 (Cheng et al. 1999). This region includes stem loop structures SL8926, SL9011 and SL906. It has yet to be elucidated whether these stem loops are specifically involved in the binding the NS5B protein, or the replication complex. However, their presence within this region and high degree of conservation between divergent genotypes is suggestive that they may be involved in strand initiation in the same way as the picornavirus CREs template strand extension from the 3' end of the genome.

The 3'UTR of HCV has been shown to be absolutely required for virus infectivity in the chimpanzee animal model (Yanagi et al. 1999). It has also been shown that the helicase domain of NS3 and polpyrimidine tract-binding protein (PTB) bind to stem loop structures within this region (Banerjee and Dasgupta 2001) (Ito and Lai 1997). NS3 forms a complex with NS5B and is required to unwind paired RNA and PTB binds both the 5' and 3' UTRs circularising the full length RNA; suggesting that RNA structures within the 3'UTR may play a role in viral replication and/or regulation of translation. The 5' boundary of the RNA structures within the 3'UTR has yet to be defined. Consequently it is possible that the binding of the helicase domain of NS3 and the PTB protein to stem loop structures in the 3'UTR (Banerjee and Dasgupta 2001; Ito and Lai 1997), may involve additional interactions with adjacent RNA structures in the NS5B region.

The stem loop structures predicted in this study would be amenable to functional investigation through mutational analysis of replicating clones of HCV. Unfortunately, the only animal model available for HCV infection and replication is currently the chimpanzee and its use is limited for ethical reasons, its scarcity and high cost of maintenance. Further more, no cell culture system for HCV has yet been described in which the virus replicates to a sufficiently titre or level of reproducibility to allow detailed *in vitro* analysis of its molecular biology.

A subgenomic replicon system, in which the structural protein coding regions have been removed, has recently been developed which replicates to high titre in the human hepatoma cell line Huh-7 (Lohmann et al. 1999). The replicon system would allow mutational analysis of the stem loop structures within the NS5B region. Unfortunately the stem loops predicted towards the 5' of the coding region fall within the core gene which has been removed from the replicon. The lack of structural region may also disrupt potential long range RNA-RNA interactions making the system less realistic. It would be possible to investigate potential interactions between the RNA structures and viral or cellular proteins with a *Saccharomyses cerevisiae* three-hybrid system (SenGupta et al. 1996). This has recently been used to study the interactions between the 3'X region of HCV and human ribosomal proteins (Wood et al. 2001).

In summary, ribonuclease cleavage analysis has been used to physically identify the position and structure of eight stable stem loop structures within the HCV coding region. The role of these stem loops has yet to be established but evidence presented in this study of their conservation between diverse genotypes, close proximity to the highly structured 5' and 3' UTRs and possible long range RNA-RNA interactions is

consistent with a role in one or more aspects of virus replication and/or interaction with the cell.

CHAPTER 7

FINAL DISCUSSION

7. FINAL DISCUSSION

Recently a number of observations have suggested the existence of constraints on the sequence divergence of HCV, HGV/GBV-C and related viruses. Various lines of evidence, including the geographical spread of the virus and infection of old and new world primates, suggest that HGV/GBV-C has always infected the human population through co-speciation. However, the lack of sequence diversity and high rate of nucleotide substitution observed in the virus suggests a relatively recent time of infection and diversity in the human population. This apparent contradiction has been accounted for with the observation of reduced variability at synonymous sites and clustering of covariant substitutions which suggest constraints on sequence divergence, out with those normally imposed by the requirement to encode functional protein translation products. It has been proposed that such constraints are due to the presence of evolutionarily conserved RNA structure within the coding region of the virus genome. The genome of HCV exhibits greater diversity than HGV/GBV-C, although constraints on sequence change at synonymous sites and clustering of covariant substitutions are observed, in particular towards the 5' and 3' regions of the genome. It has been suggested that these biases in sequence divergence are due to the presence of evolutionarily conserved RNA structure, in which the requirement to maintain nucleotide pairing constrains the potential for nucleotide substitutions.

In order to fully investigate potential RNA structure in the coding regions of HCV and HGV/GBV-C the genomes of both viruses were analysed for regions of excess folding free energy (FFEs), which was shown as being indicative of sequence

dependant RNA structure (chapter 3). Previously published analysis of RNA structure in eukaryotic and prokaryotic genomes reveals a number of potential causes of artefactual results such as the disruption of regional differences in nucleotide and dinucleotide composition in the calculation of FFE differences, between native and randomised sequences. In order to overcome this, a number of sequence randomisation methods where developed and assessed, which maintain local sequence heterogeneity during analysis. Each method gave comparable results, indicating that figures of FFE presented in this study are not due to the homogenisation of composition biases within the sequences analysed.

Large excesses in FFE were observed across the genomes of both HCV and HGV/GBV-C. The results were comparable to those of known highly structured RNA genomes, such as plant viroids and the non-coding region of hepatitis delta virus and are suggestive of extensive sequence dependent RNA structure across the coding regions of HCV and HGV/GBV-C. The greatest FFE differences were observed within the core gene and NS5B coding region of HCV and NS5B region of HGV/GBV-C which is in close concordance with regions of reduced synonymous variability and clustering of covariant substitutions.

These observations of large differences in FFE across the coding regions of HCV and HGV/GBV-C suggests that sequence dependent RNA structure is distributed throughout the virus genomes and is not restricted to the 5' and 3' UTRs. Such structure would account for the biases observed in nucleotide substitutions within HCV and HGV/GBV-V and the lack of sequence diversity observed in HGV/GBV-C and related viruses.

Specific predictions of RNA secondary structure were made for the NS5B regions of both viruses as well as the core gene of HCV (chapter 4). Predictions were made using a combination of thermodynamic and phylogenetic methods including free energy minimisation on folding, structure conservation between genotypes and the occurrence of both covariant and semi-covariant substitutions. Combining these methods, the computational data presented in this thesis demonstrates and predicts the existence and structure of at least ten evolutionarily conserved stem loops in the core gene and NS5B region of HCV (Tuplin et al., 2002). The structure of a further eight stem loops was demonstrated in the NS5B region and seven in the 3'UTR of HGV/GBV-C (Cuceanu et al., 2001). High frequencies of covariant and semi-covariant substitutions were noted within all but the terminal two structures within the 3'UTR of HGV/GBV-C, in which extreme conservation of nucleotide sequence was observed.

These predictions confirm the existence of a number of structures predicted independently by simple sequence inspection (Han and Houghton, 1992; Smith and Simmonds, 1997), or by a comparative RNA-folding algorithm that identified covariant sites through the comparison of phylogenetically conserved RNA structures (Hofacker et al., 1998). The former two studies predicted the existence of SL9011 and SL9118; the latter SL7729 and the terminal seven out of twenty four paired nucleotides in SL8828. However, with the exception of SL9118, there are a number of discrepancies between the conformation of previously predicted structures and those presented in this thesis. These differences may be explained by the expanded data set used in this study, which included all epidemiologically unlinked HCV and HGV/GBV-C complete genome sequences available on GenBank, and the

shorter distances over which the sequences were previously analysed which may have limited the potential for RNA folding.

Free energy minimisation and RNase cleavage experiments have previously demonstrated the existence of stable secondary structure towards the distal part of the 3'UTR of number of different flaviviruses (Proutski et al., 1997; Brinton et al., 1986; Rice et al., 1985), HCV (Kolykhalov et al., 1996; Tanaka et al., 1996; Blight and Rice, 1997), GBV-A and GBV-B (Sbardellati et al., 1999), pestiviruses (Yu et al., 1999; Deng and Brock, 1993) and the *Picornaviridae* family (Witwer et al., 2001; Mellits et al., 1998). Further studies have provided evidence of specific interactions between these regions and both viral and cellular proteins involved in virus replication (Banerjee and Dasgupta, 2001; Ito and Lai, 1997) and determined which regions of the HCV 3'UTR are critical for replication of HCV in a chimpanzee model (Yanagi et al., 1999). The terminal two structures predicted in this study in the extreme down-stream region of the HGV/GBV-C 3'UTR closely resemble those predicted in HCV, GBV-A and GBV-B. However, there is no evidence for a conserved third stem loop in HGV/GBV-C which forms a characteristic cloverleaf conformation in HCV, GBV-A and GBV-B. Further, the terminal loop of HGV/GBV-C is shorter (14 nucleotide pairs) than those previously predicted in HCV (19 nucleotide pairs) and GBV-B (21 nucleotide pairs).

A multisite RNA transcript priming method was used to enzymatically map the extensive predicted RNA structures in the core gene and NS5B region of HCV (Chapter 6). The use of long RNA transcripts for nuclease digestion was essential to allow stem loop formation to occur in as natural a context as possible, and in particular to enable potential interactions between RNA secondary structures, such as

the formation of pseudoknots, to occur prior to nuclease cleavage. Whilst the nuclease mapping data generally corresponds to the predicted structures for genotypes 1a and 2a, there were some instances where the two methods make discordant predictions. Such discrepancies included the observation of V_1 (paired nucleotides) cleavage at one or two out of the five bases in the single stranded terminal loop of SL47 and the predicted single stranded bulge within the base paired stem. This and similar observations, including apparently both single stranded and paired residues, in a number of stem loops provide evidence for additional base pairing; suggesting the formation of higher order structures such as pseudoknots between adjacent secondary structures.

Experimental data on the function of RNA structures in the genome of RNA viruses is far from complete, particularly for those in the genomes of HCV and HGV/GBV-C described here. However, evidence for a role of RNA structure within the core gene of HCV in translation has been provided by the observation of enhanced translation of a dicistronic reporter system, when a predicted long range interaction between nucleotides -317 to -302 was disrupted (Kim et al., 2003). It has also been shown that a deletion between nucleotides -317 to -302 of the 5'UTR enhances translation but only when nucleotides 66-588 of the core gene are included in the construct (Honda et al., 1999). Three RNA structures SL47, SL87 and SL443 fall within this region and may thus play functional roles in regulation of translation.

At the other end of the genome and perhaps functionally analogous to the CRE structures found in the coding regions of picornaviruses, the NS5B protein (RdRp) of HCV has been shown to preferentially bind to the NS5B RNA sequence 8922-9121, a region which includes SL8933, SL9017 and SL9067. Consequently these

structures may be involved in strand initiation in a similar way to picornavirus CRE initiated strand extension from the 3' end of the genome. The upstream limit of the functional structures mapped in the 3'UTR of HCV has yet to be fully defined. Therefore, it is possible that the binding of the helicase domain of NS3 and the PTB protein to stem loop structures in the 3'UTR may involve further interactions with adjacent structures within the NS5B region.

Negative staining followed by transmission electron microscopy and gold label hybridisation was used to directly visualise the folding conformation of the NS5B coding region and 3'UTR of HGV/GBV-C (chapter 5). The NS5B coding region was observed to fold onto a cruciform conformation which was consistent with computational predictions of overall RNA folding conformation, based on free energy minimisation. Even though electron microscopy provided information on the identity and presence of a number of stem loop structures, within the NS5B coding region of HGV/GBV-C, it proved less useful than ribonuclease mapping in providing detailed information on discrete stem loop structures. In particular definitive visualisation of the 3'UTR of HGV/GBV-C was not successful. This may have been due to the fact that less RNA folding and shorter structures were predicted in this region, which were not large enough to resolve against the negatively stained background.

Electron microscopy proved more useful in visualisation of the genome wide RNA folding structure of HGV/GBV-C; predicted to exist by excess FFE in HCV and to a greater extent HGV/GBV-C. The genome of HGV/GBV-C was observed to fold into four main negatively stained structures radiating from a central core. The IRES (as confirmed by specific gold label hybridisation) was presented at a kink

approximately half was along a radiating stem. The observed genome wide folding of HGV/GBV-C was similar to that previously observed of HCV (Lucy Beales, personal communication).

In summary, a combination of thermodynamic, phylogenetic and physical methods have been used in this thesis to computationally and physically identify and confirm a number of conserved RNA structures in the coding regions of both HCV and HGV/GBV-C. Excess folding free energies have also indicated that genome wide RNA folding may exist in both viruses, although it is not clear why this might be more extensive for HGV/GBV-C than HCV. However the degree of structural conservation between diverse genotypes is consistent with roles for the predicted structures in one or more aspects of virus replication and/or interaction with the cell. Whilst these are currently not understood in any detail, the downstream HCV structures predicted in this thesis are providing a starting point for such functional studies using HCV replicons.

REFERENCES

Aach,R.D., Stevens,C.E., Hollinger,F.B., Mosley,J.W., Peterson,D.A., Taylor,P.E., Johnson,R.G., Barbosa,L.H., and Nemo,G.J. (1991). Hepatitis C virus infection in post-transfusion hepatitis. An analysis with first- and second-generation assays. N. Engl. J. Med. *325*, 1325-1329.

Abe,K., Inchauspe,G., and Fujisawa,K. (1992). Genomic characterisation and mutation rate of hepatitis C virus isolated from a patient who contracted hepatitis during an epidemic on non-A, non-B hepatitis in Japan. J. Gen. Virol. *73*, 2725-2729.

Adams, N. J., Prescott, L. E., Jarvis, L. M., Lewis, J. C. M., Mcclure, M. O., Smith, D. B., and Simmonds, P. (1998). Detection of a novel flavivirus related to hepatitis G virus/GB virus C in chimpanzees. J. Gen. Virol. 79, 1871-1877. 1998.

Almeida,J.D., Deinhardt,F., Holmes,A.W., Peterson,D.A., Wolfe,L., and Zuckerman,A.J. (1976). Morphology of the GB hepatitis agent. Nature *261*, 608-609.

Alter,H.J., Holland,P.V., Morrow,A.G., Purcell,R.H., Feinstone,S.M., and Moritsugu,Y. (1975). Clinical and serological analysis of transfusion-associated hepatitis. Lancet *ii*, 838-841.

Alter, H. J., Nakatsuji, Y., Melpolder, J., Wages, J., Wesley, R., Shih, J. W. K., and Kim, J. P. (1997a). The incidence of transfusion-associated hepatitis G virus infection and its relation to liver disease. N. Engl. J. Med. 336[11], 747-754.

Alter,H.J., Purcell,R.H., Holland,P.V., and Popper,H. (1978). Transmissible agent in non-A, non-B hepatitis. Lancet *1*, 459-463.

Alter,H.J., Purcell,R.H., Shih,J.W., Melpolder,J.C., Houghton,M., Choo,Q.L., and Kuo,G. (1989). Detection of antibody to hepatitis C virus in prospectively followed transfusion recipients with acute and chronic non-A, non- B hepatitis. N. Engl. J. Med. *321*, 1494-1500.

Alter,M.J. (1997c). Epidemiology of hepatitis C. Hepatology *26*, S62-S65.

Alter, M. J., Gallagher, M., Morris, T. T., Moyer, L. A., Meeks, E. L., Krawczynski, K., Kim, J. P., and Margolis, H. S. (1997b). Acute non-A-E hepatitis in the United States and the role of hepatitis G virus infection. N. Engl. J. Med. 336[11], 741-746.

Alter, M. J., KruszonMoran, D., Nainan, O. V., McQuillan, G. M., Gao, F. X., Moyer, L. A., Kaslow, R. A., and Margolis, H. S. (1999). The prevalence of hepatitis C virus infection in the United States, 1988 through 1994. N Engl J Med 341[8], 556-562.

Alter,M.J., Margolis,H.S., Krawczynski,K., Judson,F.N., Mares,A., Alexander,W.J., Hu,P.Y., Miller,J.K., Gerber,M.A., Sampliner,R.E., Meeks,E.L., and Beach,M.J. (1992). The natural history of community-acquired hepatitis C in the united states. N. Engl. J. Med. *327*, 1899-1905.

Aymard,J.P., Botte,C., Contal,P., Janot,C., and Streiff,F. (1993). Seroprevalence of hepatitis C antibodies among blood donors - a study of 2nd generation ELISA and RIBA tests and surrogate markers. Pathol. Biol. *41*, 149-153.

Bain,C., Fatmi,A., Zoulim,F., Zarski,J.P., Trepo,C., and Inchauspe,G. (2001). Impaired allostimulatory function of dendritic cells in chronic hepatitis C infection. Gastroenterology *120*, 512-524.

Ballardini,G., Groff,P., Pontisso,P., Giostra,F., Francesconi,R., Lenzi,M., Zauli,D., Alberti,A., and Bianchi,F.B. (1995). Hepatitis C virus (HCV) genotype, tissue HCV antigens, hepatocellular expression of HLA-a,b,c, and intercellular adhesion-1 molecules - clues to pathogenesis of hepatocellular damage and response to interferon treatment in patients with chronic hepatitis C. J. Clin. Invest. *95*, 2067-2075.

Banerjee,R. and Dasgupta,A. (2001). Specific interaction of hepatitis C virus protease/helicase NS3 with the 3'-terminal sequences of viral positive- and negative-strand RNA. J. Virol. *75*, 1708-21.

Bartenschlager,R., Ahlbornlaake,L., Mous,J., and Jacobsen,H. (1993). Nonstructural protein-3 of the hepatitis C virus encodes a serine-type proteinase required for cleavage at the NS3/4 and NS4/5 junctions. J. Virol. *67*, 3835-3844.

Bassit, L., Kleter, B., RibeirodosSantos, G., Maertens, G., Sabino, E., Chamone, D., Quint, W., and SaezAlquezar, A. (1998). Hepatitis G virus: Prevalence and sequence analysis in blood donors of Sao Paulo, Brazil. Vox Sang. 74[2], 83-87.

Beales,L.P., Rowlands,D.J., and Holzenburg,A. (2001). The internal ribosome entry site (IRES) of hepatitis C virus visualized by electron microscopy. RNA *7*, 661-670.

Beales,L.P., Holzenburg,A., and Rowlands,D.J. (2003). Viral internal ribosome entry site structures segregate into two distinct morphologies. J. Virol. *77*, 6574-6579.

Belyaev, A. S., Chong, S., Novikov, A., Kongpachith, A., Masiarz, F. R., Lim, M., and Kim, J. P. (1998). Hepatitis G virus encodes protease activities which can effect processing of the virus putative nonstructural proteins. J. Virol. 72[1], 868-872.

Benezra,J. (1993). Sexual transmission of HCV. Lancet *342*, 626.

Benvegnu, L. B., Pontisso, P., Cavalletto, D., Noventa, F., Chemello, L., and Alberti, A. Lack of correlation between hepatitis C virus genotypes and clinical course of hepatitis C virus-related cirrhosis. (1997). Hepatology 25[1], 211-215.

Bernvil,S.S., Andrews,V.J., and Sasich,F. (1994). Second-generation anti-HCV screening in a Saudi Arabian donor population. Vox Sang. *66*, 33-36

Blackwell,J.L. and Brinton,M.A. (1995). BHK cell proteins that bind to the 3' stem-loop structure of the West Nile virus genome RNA. J. Virol. *69*, 5650-5658.

Blight,K.J. and Rice,C.M. (1997). Secondary structure determination of the conserved 98-base sequence at the 3' terminus of hepatitis C virus genome RNA. J. Virol. *71*, 7345-7352.

Bosmans, J. L., Nouwen, E. J., Behets, G., Gorteman, K., Huraib, S. O., Shaheen, F. A., Maertens, G., Verpooten, G. A., Elseviers, M. M., deBroe, M. E., Billiouw, J. M., Cosyn, L., Daelemans, R., Moeremans, C., VanRoost, G., VanderNiepen, P., Schurgers, M., Donck, J., Segaert, M., Verbanck, J., Bosteels, V., Hombrouckx, R., and DePaepe, M. (1997). Prevalence and clinical expression of HCV-genotypes in haemodialysis-patients of two geographically remote countries: Belgium and Saudi-Arabia. Clin.Nephrol. 47[4], 256-262.

Bradley, D. W., Maynard, J. E., Popper, H., Cook, E. H., Ebert, J. W., McCaustland, K. A., Schable, C. A., and Fields, H. A. (1983). Post-transfusion non-A, non-B hepatitis: physicochemical properties of two distinct agents. J. Inf. Dis. 148, 254-265.

Bradley,D.W., McCaustland,K.A., Cook,E.H., Schable,C.A., Ebert,J.W., and Maynard,J.E. (1985). Posttransfusion non-A, non-B hepatitis in chimpanzees. Physicochemical evidence that the tubule-forming agent is a small, enveloped virus. Gastroenterology *88*, 773-779.

Bralet, M. P., Roudotthoraval, F., Pawlotsky, J. M., Bastie, A., Vannhieu, J. T., Duval, J., Dhumeaux, D., and Zafrani, E. S. (1997). Histopathologic impact of GB virus C infection on chronic hepatitis C. Gastroenterology 112[1], 188-192.

Branch,A.D. and Robertson,H.D. (1984). A replication cycle for viroids and other small infectious RNA's. Science *223*, 450-455.

Bresters,D., Mauserbunschoten,E.P., Reesink,H.W., Roosendaal,G., van der Poel,C.L., Chamuleau,R.A.F.M., Jansen,P.L.M., Weegink,C.J., Cuypers,H.T.M.,

Lelie,P.N., and Vandenberg,H.M. (1993). Sexual transmission of hepatitis C virus. Lancet *342*, 210-211.

Brinton,M.A., Fernandez,A.V., and Dispoto,J.H. (1986). The 3'-nucleotides of flavivirus genomic RNA form a conserved secondary structure. Virology *153*, 113-121.

Brown,E.A., Zhang,H.C., Ping,L.H., and Lemon,S.M. (1992). Secondary structure of the 5' nontranslated regions of hepatitis C virus and pestivirus genomic RNAs. Nucleic. Acids. Res. *20*, 5041-5045.

Bukh,J., Purcell,R.H., and Miller,R.H. (1992). Sequence analysis of the 5' noncoding region of hepatitis C virus. Proc. Natl. Acad. Sci. USA *89*, 4942-4946.

Chamberlain, R. W., Adams, N., Saeed, A. A., Simmonds, P., and Elliott, R. M. Complete nucleotide sequence of a type 4 hepatitis C virus variant, the predominant genotype in the Middle East. (1997). J. Gen. Virol. 78, 1341-1347.

Chang, K. M., Rehermann, B., McHutchison, J. G., Pasquinelli, C., Southwood, S., Sette, A., and Chisari, F. V. Immunological significance of cytotoxic T lymphocyte epitope variants in patients chronically infected by the hepatitis C virus. (1997). J. Clin. Invest. 100[9], 2376-2385.

Charrel, R. N., de Micco, P., and de Lamballerie, X. Phylogenetic analysis of GB viruses A and C: evidence for cospeciation between virus isolates and their primate hosts. (1999). J. Gen. Virol. 80, 2329-2335.

Chen, H. L., Chang, M. H., Ni, Y. H., Hsu, H. Y., Kao, J. H., and Chen, P. J. Hepatitis G virus infection in normal and prospectively followed posttransfusion children. (1997). Pediatr Res 42[6], 784-787.

Cheng, J. C., Chang, M. F., and Chang, S. C. Specific interaction between the hepatitis C virus NS5B RNA polymerase and the 3 ' end of the viral RNA. (1999). J Virol 73[8], 7044-7049.

Chiaramonte, M., Stroffolini, T., Vian, A., Stazi, M. A., Floreani, A., Lorenzoni, U., Lobello, S., Farinati, F., and Naccarato, R. (1999). Rate of incidence of hepatocellular carcinoma in patients with compensated viral cirrhosis. Cancer 85[10], 2132-2137.

Choo,Q.L., Kuo,G., Weiner,A.J., Overby,L.R., Bradley,D.W., and Houghton,M. (1989). Isolation of a cDNA derived from a blood-borne non-A, non-B hepatitis genome. Science *244*, 359-362.

Choo,Q.L., Richman,K.H., Han,J.H., Berger,K., Lee,C., Dong,C., Gallegos,C., Coit,D., Medina Selby,R., Barr,P.J., Weiner,A.J., Bradley,D.W., Kuo,G., and Houghton,M. (1991). Genetic organization and diversity of the hepatitis C virus. Proc. Natl. Acad. Sci. USA *88*, 2451-2455.

Clarke,B. (1997). Molecular virology of hepatitis C virus. J. Gen. Virol. *78:2397-2410*, 2397-2410.

Crofts, N., Jolley, D., Kaldor, J., Vanbeek, I., and Wodak, A. (1997). Epidemiology of hepatitis C virus infection among injecting drug users in Australia. J.Epidemiol.Community.Health 51[6], 692-697.

Cuceanu,N.M., Tuplin,A., and Simmonds,P. (2001). Evolutionarily conserved RNA secondary structures in coding and non-coding sequences at the 3' end of the hepatitis G virus/GB- virus C genome. J. Gen. Virol. *82*, 713-22.

Deinhardt,F., Holmes,A.W., Capps,R.B., and Popper,H. (1967). Studies on the transmission of human viral hepatitis to marmoset monkeys. I. Transmission of disease, serial passages, and description of liver lesions. J. Exp. Med. *125*, 673-688.

Deka,N., Sharma,M.D., and Mukerjee,R. (1994). Isolation of the novel agent from human stool samples that is associated with sporadic non-A, non-B hepatitis. J. Virol. *68*, 7810-7815.

Deleersnyder, V., Pillez, A., Wychowski, C., Blight, K., Xu, J., Hahn, Y. S., Rice, C. M., and Dubuisson, J. (1997). Formation of native hepatitis C virus glycoprotein complexes. J. Virol. 71[1], 697-704.

Deng,R.T. and Brock,K.V. (1993). 5' and 3' untranslated regions of pestivirus genome - primary and secondary structure analyses. Nucleic. Acids. Res. *21*, 1949-1957.

Di Bisceglie,A.M. (1998). Hepatitis C. Lancet *351*, 351-355.

Diepolder, H. M., Gerlach, J. T., Zachoval, R., Hoffmann, R. M., Jung, M. C., Wierenga, E. A., Scholz, S., Santantonio, T., Houghton, M., Southwood, S., Sette, A., and Pape, G. R. (1997). Immunodominant CD4(+) T-cell epitope within nonstructural protein 3 in acute hepatitis C virus infection. J. Virol. 71[8], 6011-6019.

Diepolder,H.M., Zachoval,R., Hoffmann,R.M., Wierenga,E.A., Santantonio,T., Jung,M.C., Eichenlaub,D., and Pape,G.R. (1995). Possible mechanism involving t-lymphocyte response to non-structural protein 3 in viral clearance in acute Hepatitis C virus infection. Lancet *346*, 1006-1007.

Dock-Bregeon,A.C., Chevrier,B., Podjarny,A., Johnson,J., de Bear,J.S., Gough,G.R., Gilham,P.T., and Moras,D. (1989). Crystallographic structure of an RNA helix: [U(UA)6A]2. J. Mol. Biol. *209*, 459-474.

Enomoto,N., Sakuma,I., Asahina,Y., Kurosaki,M., Murakami,T., Yamamoto,C., Izumi,N., Marumo,F., and Sato,C. (1995). Comparison of full-length sequences of interferon- sensitive and resistant hepatitis C virus 1b - sensitivity to interferon is conferred by amino acid substitutions in the NS5a region. J. Clin. Invest. *96*, 224-230.

Failla,C., Tomei,L., and Defrancesco,R. (1995). An amino-terminal domain of the hepatitis C virus NS3 protease is essential for interaction with NS4A. J. Virol. *69*, 1769-1777.

Fan, X. F., Xu, Y. J., Solomon, H., Ramrakhiani, S., NeuschwanderTetri, B. A., and Dibisceglie, A. M. (1999). Is hepatitis G/GB virus-C virus hepatotropic? Detection of hepatitis G/GB virus-C viral RNA in liver and serum. J Med Virol 58[2], 160-164.

Farci,P., Alter,H.J., Govindarajan,S., Wong,D.C., Engle,R., Lesniewski,R.R., Mushahwar,I.K., Desai,S.M., Miller,R.H., Ogata,N., and Purcell,R.H. (1992). Lack of protective immunity against reinfection with hepatitis C virus. Science *258*, 135-140.

Farci,P., Alter,H.J., Wong,D., Miller,R.H., Shih,J.W., Jett,B., and Purcell,R.H. (1991). A long-term study of hepatitis C virus replication in non-A, non- B hepatitis. N. Engl. J. Med. *325*, 98-104.

Farci,P. and Purcell,R.H. (2000). Clinical significance of hepatitis C virus genotypes and quasispecies. Semin. Liver Dis. *20*, 103-126.

Farci, P., Shimoda, A., Wong, D., Cabezon, T., De Gioannis, D., Strazzera, A., Shimizu, Y., Shapiro, M., Alter, H. J., and Purcell, R. H. (1996). Prevention of hepatitis C virus infection in chimpanzees by hyperimmune serum against the hypervariable region 1 of the envelope 2 protein. Proc.Natl.Acad.Sci.USA 93, 15394-15399.

Feinstone,S.M., Kapikian,A.Z., and Purcell,R.H. (1975). Transfusion-associated hepatitis not due to viral hepatitis A or B. N. Engl. J. Med. *292*, 767-770.

Feinstone,S.M., Mihalik,K.B., Kamimura,T., Alter,H.J., London,W.T., and Purcell,R.H. (1983). Inactivation of hepatitis B virus and non-A, non-B hepatitis by chloroform. Infect. Immun. *41*, 816-821.

Ferrero,S., Lungaro,P., Bruzzone,B.M., Gotta,C., Bentivoglio,G., and Ragni,N. (2003). Prospective study of mother-to-infant transmission of hepatitis C virus: a 10-year survey (1990-2000). Acta Obstet. Gynecol. Scand. *82*, 229-234.

Ferri, C., Lacivita, L., and Zignego, A. L. (1996). Extrahepatic manifestations of hepatitis C virus infection. Ann.Intern.Med. 125[4], 344.

Feucht,H., Zollner,B., Polywka,S., Knodler,B., Schroter,M., Nolte,H., and Laufs,R. (1997). Distribution of hepatitis G viraemia and antibody response to recombinant proteins with special regard to risk factors in 709 patients. Hepatology *26, no.2*, 491-494.

Feucht, H. H., Zollner, B., Polywka, S., and Laufs, R. (1996). Vertical transmission of hepatitis G. Lancet 347, 615-616.

Flint, M., Maidens, C., LoomisPrice, L. D., Shotton, C., Dubuisson, J., Monk, P., Higginbottom, A., Levy, S., and McKeating, J. A. (1996). Characterization of

hepatitis C virus E2 glycoprotein interaction with a putative cellular receptor, CD81. J Virol 73[8], 6235-6244. 1999.

Flint, M. and McKeating, J. A. (1999). The C-terminal region of the hepatitis C virus E1 glycoprotein confers localization within the endoplasmic reticulum. J Gen Virol 80, 1943-1947.

Frank,C., Mohamed,M.K., Strickland,G.T., Lavanchy,D., Arthur,R.R., Magder,L.S., El Khoby,T., Abdel-Wahab,Y., Aly Ohn,E.S., Anwar,W., and Sallam,I. (2000). The role of parenteral antischistosomal therapy in the spread of hepatitis C virus in Egypt. Lancet *355*, 887-891.

Fretz, C., Jeannel, D., Stuyver, L., Herve, V., Lunel, F., Boudifa, A., Mathiot, C., deThe, G., and Fournel, J. J. (1995). HCV infection in a rural population of the Central African Republic (CAR): Evidence for three additional subtypes of genotype 4. J.Med Virol. 47[4], 435-437.

Fukuda, M., Chayama, K., Tsubota, A., Kobayashi, M., Hashimoto, M., Miyano, Y., Koike, H., Koida, I., Arase, Y., Saitoh, S., Murashima, N., Ikeda, K., and Kumada, H. (1998). Predictive factors in eradicating hepatitis C virus using a relatively small dose of interferon. J.Gastroenterol.Hepatol. 13[4], 412-418.

Gabashvili,I.S., Agrawal,R.K., Grassucci,R., and Frank,J. (1999). Structure and structural variations of the Escherichia coli 30 S ribosomal subunit as revealed by three-dimensional cryo-electron microscopy. J. Mol. Biol. *286*, 1285-1291.

Gale, M., Blakely, C. M., Kwieciszewski, B., Tan, S. L., Dossett, M., Tang, N. M., Korth, M. J., Polyak, S. J., Gretch, D. R., and Katze, M. G. (1998). Control of PKR protein kinase by hepatitis C virus nonstructural 5A protein: Molecular mechanisms of kinase regulation. Mol.Cell Biol. 18[9], 5208-5218.

Gale, M. J., Korth, M. J., Tang, N. M., Tan, S. L., Hopkins, D. A., Dever, T. E., Polyak, S. J., Gretch, D. R., and Katze, M. G. (1997). Evidence that hepatitis C virus resistance to interferon is mediated through repression of the PKR protein kinase by the nonstructural 5A protein. Virology 230[2], 217-227.

GonzalezPerez, M. A., Norder, H., Bergstrom, A., Lopez, E., Visona, K. A., and Magnius, L. O. (1997). High prevalence of GB virus C strains genetically related to strains with Asian origin in Nicaraguan hemophiliacs. J. Med. Virol. 52[2], 149-155.

Goodfellow,I., Chaudhry,Y., Richardson,A., Meredith,J., Almond,J.W., Barclay,W., and Evans,D.J. (2000). Identification of a cis-acting replication element within the poliovirus coding region. J. Virol. *74* , 4590-4600.

Grakoui,A., Mccourt,D.W., Wychowski,C., Feinstone,S.M., and Rice,C.M. (1993a). Characterization of the hepatitis C virus-encoded serine proteinase - determination of proteinase-dependent polyprotein cleavage sites. J. Virol. *67*, 2832-2843.

Grakoui,A., Wychowski,C., Lin,C., Feinstone,S.M., and Rice,C.M. (1993b). Expression and identification of hepatitis C virus polyprotein cleavage products. J. Virol. *67*, 1385-1395.

Haagsma, E. B., Cuypers, H. T. M., Gouw, A. S. H., Sjerps, M. C., Huizenga, J. R., Slooff, M. J. H., and Jansen, P. L. M. (1997). High prevalence of hepatitis G virus after liver transplantation without apparent influence on long-term graft function. Journal of Hepatology 26[4], 921-925.

Hammel,P., Marcellin,P., Martinotpeignoux,M., Pham,B.N., Degott,C., Level,R., Lefort,V., Benhallem,A., Erlinger,S., and Benhamou,J.P. (1994). Etiology of chronic hepatitis in france: predominant role of hepatitis C virus. J. Hepatol. *21*, 618-623.

Han,J.H. and Houghton,M. (1992). Group specific sequences and conserved secondary structures at the 3' end of HCV genome and its implication for viral replication. Nucleic. Acids. Res. *20*, 3520.

Han,J.H., Shyamala,V., Richman,K.H., Brauer,M.J., Irvine,B., Urdea,M.S., Tekamp Olson,P., Kuo,G., Choo,Q.L., and Houghton,M. (1991). Characterization of the terminal regions of hepatitis C viral RNA: identification of conserved sequences in the 5' untranslated region and poly(A) tails at the 3' end. Proc. Natl. Acad. Sci. USA *88*, 1711-1715.

Hara, T., Setoguchi, Y., Kajihara, S., Yamamoto, K., Sakai, T., Inoue, T., Ohba, K., and Mizokami, M. (1996). Phylogenetic tree-based epidemiological analysis of hepatitis C virus transmission in a region of Japan with a high prevalence of infection. J.Gastroenterol.Hepatol. 11[7], 641-645.

Harris,K.A., Gilham,C., Mortimer,P.P., and Teo,C.G. (1999). The most prevalent hepatitis C virus genotypes in England and Wales are 3a and 1a. J. Med. Virol. *58*, 127-131.

Heringlake,S., Osterkamp,S., Trautwein,C., Tillmann,H.L., Boker,K., Muerhoff,S., Mushahwar,I.K., Hunsmann,G., and Manns,M.P. (1996). Association between fulminant hepatic failure and a strain of GBV virus C. Lancet *348*, 1626-1629.

Hijikata,M., Kato,N., Ootsuyama,Y., Nakagawa,M., Ohkoshi,S., and Shimotohno,K. (1991a). Hypervariable regions in the putative glycoprotein of hepatitis C virus. Biochem. Biophys. Res. Commun. *175*, 220-228.

Hijikata,M., Kato,N., Ootsuyama,Y., Nakagawa,M., and Shimotohno,K. (1991b). Gene mapping of the putative structural region of the hepatitis C virus genome by in vitro processing analysis. Proc. Natl. Acad. Sci. USA *88*, 5547-5551.

Hijikata,M., Mizushima,H., Akagi,T., Mori,S., Kakiuchi,N., Kato,N., Tanaka,T., Kimura,K., and Shimotohno,K. (1993). 2 distinct proteinase activities required for the processing of a putative nonstructural precursor protein of hepatitis C virus. J. Virol. *67*, 4665-4675.

Hofacker,I.L., Fekete,M., Flamm,C., Huynen,M.A., Rauscher,S., Stolorz,P.E., and Stadler,P.F. (1998). Automatic detection of conserved RNA structure elements in complete RNA virus genomes. Nucleic Acids Res. *26*, 3825-3836.

Honda, M., Brown, E. A., and Lemon, S. M. (1996a). Stability of a stem-loop involving the initiator AUG controls the efficiency of internal initiation of translation on hepatitis C virus RNA. RNA 2[10], 955-968.

Honda, M., Ping, L. H., Rijnbrand, R. C. A., Amphlett, E., Clarke, B., Rowlands, D., and Lemon, S. M. (1996b). Structural requirements for initiation of translation by internal ribosome entry within genome-length hepatitis C virus RNA. Virology 222[1], 31-42.

Honda, M., Rijnbrand, R., Abell, G., Kim, D. S., and Lemon, S. M. (1999). Natural variation in translational activities of the 5 ' nontranslated RNAs of hepatitis C virus genotypes 1a and 1b: Evidence for a long-range RNA-RNA interaction outside of the internal ribosomal entry site. J Virol 73[6], 4941-4951.

Hope,V.D., Judd,A., Hickman,M., Lamagni,T., Hunter,G., Stimson,G.V., Jones,S., Donovan,L., Parry,J.V., and Gill,O.N. (2001). Prevalence of hepatitis C among injection drug users in England and Wales: is harm reduction working? Am J. Public Health *91*, 38-42.

Hyland, C. A., Mison, L., Solomon, N., Cockerill, J., Wang, L., Hunt, J., Selvey, L. A., Faoagali, J., Cooksley, W. G. E., Young, I. F., Trowbridge, R., Borthwick, I., and Gowans, E. J. (1998). Exposure to GB virus type C or hepatitis G virus in selected Australian adult and children populations. Transfusion 38[9], 821-827.

Ina,Y., Mizokami,M., Ohba,K., and Gojobori,T. (1994). Reduction of synonymous substitutions in the core protein gene of hepatitis C virus. J. Mol. Evol. *38*, 50-56.

Ito,T. and Lai,M.M.C. (1997). Determination of the secondary structure of and cellular protein binding to the 3'-untranslated region of the hepatitis C virus RNA genome. J. Virol. *71*, 8698-8706.

Ito, T., Tahara, S. M., and Lai, M. M. C. (1998). The 3'-untranslated region of hepatitis C virus RNA enhances translation from an internal ribosomal entry site. J. Virol. 72[11], 8789-8796.

Jarvis, L. M., Davidson, F., Hanley, J. P., Yap, P. L., Ludlam, C. A., and Simmonds, P. (1996). Infection with hepatitis G virus among recipients of plasma products. Lancet 348[9038], 1352-1355.

Jin,L. and Peterson,D.L. (1995). Expression, isolation, and characterization of the hepatitis C virus ATPase/RNA helicase. Accanemic Press, Inc 6144, 6-25.

Johnson,R.J., Gretch,D.R., Yamabe,H., Hart,J., Bacchi,C.E., Hartwell,P., Couser,W.G., Corey,L., Wener,M.H., Alpers,C.E., and Willson,R. (1993). Membranoproliferative glomerulonephritis associated with hepatitis C virus infection. N. Engl. J. Med. *328*, 465-470.

Kao, J. H., Chen, W., Chen, P. J., Lai, M. Y., and Chen, D. S. (1999). Liver and peripheral blood mononuclear cells are not major sites for GB virus-C/hepatitis G virus replication. Arch Virol 144[11], 2173-2183.

Kao,J.H., Liu,C.J., Chen,P.J., Chen,W., Lai,M.Y., and Chen,D.S. (2000). Low incidence of hepatitis C virus transmission between spouses: a prospective study. J. Gastroenterol. Hepatol. *15*, 391-395.

Karayiannis,P., Petrovic,L.M., Fry,M., Moore,D., Enticott,M., McGarvey,M.J., Scheuer,P.J., and Thomas,H.C. (1989). Studies of GB hepatitis agent in tamarins. Hepatology *9*, 186-192.

Katayama, K., Fukushi, S., Kurihara, C., Ishiyama, N., Okamura, H., Hoshino, F. B., and Oya, A. (1997). New variant groups identified from HGV isolates. Arch.Virol. 142[5], 1021-1028.

Katayama, K., Kageyama, T., Fukushi, S., Hoshino, F. B., Kurihara, C., Ishiyama, N., Okamura, H., and Oya, A. (1998). Full-length GBV-C/HGV genomes from nine Japanese isolates: characterization by comparative analyses. Arch.Virol. 143[6], 1063-1075.

Kato,N., Ootsuyama,Y., Ohkoshi,S., Nakazawa,T., Sekiya,H., Hijikata,M., and Shimotohno,K. (1992). Characterization of hypervariable regions in the putative envelope protein of hepatitis C virus. Biochem. Biophys. Res. Commun. *189*, 119-127.

Kew, M. C. Hepatitis viruses and hepatocellular carcinoma. (1998). Res.Virol. 149[5], 257-262.

Khudyakov, Y. E., Cong, M. E., Bonafonte, M. T., Abdulmalek, S., Nichols, B. L., Lambert, S., Alter, M. J., and Fields, H. A. (1997). Sequence variation within a nonstructural region of hepatitis G virus genome. J. Virol. 71[9], 6875-6880.

Kim, J. L., Morgenstern, K. A., Lin, C., Fox, T., Dwyer, M. D., Landro, J. A., Chambers, S. P., Markland, W., Lepre, C. A., OMalley, E. T., Harbeson, S. L., Rice, C. M., Murcko, M. A., Caron, P. R., and Thomson, J. A. (1996). Crystal structure of the hepatitis C virus NS3 protease domain complexed with a synthetic NS4A cofactor peptide. Cell 87[2], 343-355.

Kim, Y. K., Lee, S. H., Seol, S. K., and Jang, S. K. (2003). Long-range RNA-RNA interaction between the 5' nontranslated region and the core-coding sequences of hepatitis C virus modulates the IRES-dependent translation. RNA 9, 599-606.

Kolykhalov, A. A., Feinstone, S. M., and Rice, C. M. (1996). Identification of a highly conserved sequence element at the 3' terminus of hepatitis C virus genome RNA. J. Virol. 70[6], 3363-3371.

Kondo, Y., Mizokami, M., Nakano, T., Kato, T., Ohba, K., Orito, E., Ueda, R., Mukaide, M., Hikiji, K., Oyunsuren, T., and Cooksley, W. G. (1997). Genotype of GB virus C hepatitis G virus by molecular evolutionary analysis. Virus Res. 52[2], 221-230.

Konomi,N., Miyoshi,C., La Fuente Zerain,C., Li,T.C., Arakawa,Y., and Abe,K. (1999). Epidemiology of hepatitis B, C, E, and G virus infections and molecular analysis of hepatitis G virus isolates in Bolivia. J. Clin. Microbiol. *37*, 3291-3295.

Koonin,E.V. (1991). The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses. J. Gen. Virol. *72*, 2197-2206.

Kumar,U., Brown,J., Monjardino,J., and Thomas,H.C. (1993). Sequence variation in the large envelope glycoprotein (E2/NS1) of hepatitis C virus during chronic infection. J. Infect. Dis. *167*, 726-730.

Kuo,G., Choo,Q.L., Alter,H.J., Gitnick,G.L., Redeker,A.G., Purcell,R.H., Miyamura,T., Dienstag,J.L., Alter,M.J., Stevens,C.E., Tegtmeier,F., Bonino,F., Columbo,M., Lee,W.-S., Kuo,C., Berger,K., Schuster,J.R., Overby,L.R., Bradley,D.W., and Houghton,M. (1989). An assay for circulating antibodies to a major etiologic virus of human non-A, non-B hepatitis. Science *244*, 362-364.

Laras, A., Zacharakis, G., and Hadziyannis, S. J. (1999). Absence of the negative strand of GBV-C/HGV RNA from the liver. J Hepatol 30[3], 383-388.

Laskus, T., Radkowski, M., Wang, L. F., Vargas, H., and Rakela, J. (1998). Detection of hepatitis G virus replication sites by using highly strand-specific Tth-based reverse transcriptase PCR. J. Virol. 72[4], 3072-3075.

Leary, T. P., Muerhoff, A. S., Simons, J. N., PilotMatias, T. J., Erker, J. C., Chalmers, M. L., Schlauder, G. G., Dawson, G. J., Desai, S. M., and Mushahwar, I. K. (1996). Sequence and genomic organization of GBV-C: A novel member of the flaviviridae associated with human non-A-E hepatitis. J.Med Virol. 48[1], 60-67.

Lin,C., Pragai,B.M., Grakoui,A., Xu,J., and Rice,C.M. (1994). Hepatitis C virus NS3 serine proteinase: trans-cleavage requirements and processing kinetics. J. Virol. *68*, 8147-8157.

Linnen, J., Wages, J., ZhangKeck, Z. Y., Fry, K. E., Krawczynski, K. Z., Alter, H., Koonin, E., Gallagher, M., Alter, M., Hadziyannis, S., Karayiannis, P., Fung, K., Nakatsuji, Y., Shih, J. W. K., Young, L., Piatak, M., Hoover, C., Fernandez, J., Chen, S., Zou, J. C., Morris, T., Hyams, K. C., Ismay, S., Lifson, J. D., Hess, G., Foung, S. K. H., Thomas, H., Bradley, D., Margolis, H., and Kim, J. P. (1996). Molecular cloning and disease association of hepatitis G virus: A transfusion-transmissible agent. Science 271[5248], 505-508.

Lisitsyn,N., Lisitsyn,N., and Wigler,M. (1993). Cloning the differences between two complex genomes. Science *259*, 946-951

Lobert,P.E., Escriou,N., Ruelle,J., and Michiels,T. (1999). A coding RNA sequence acts as a replication signal in cardioviruses. Proc. Natl. Acad. Sci. U. S. A. *96*, 11560-5.

Lohmann, V., Korner, F., Koch, J. O., Herian, U., Theilmann, L., and Bartenschlager, R. (1999). Replication of subgenomic hepatitis C virus RNAs in a hepatoma cell line. Science 285[5424], 110-113.

Lu, H. H. and Wimmer, E. (1996). Poliovirus chimeras replicating under the translational control of genetic elements of hepatitis C virus reveal unusual properties of the internal ribosomal entry site of hepatitis C virus. Proc.Natl.Acad.Sci.USA 93[4], 1412-1417.

Lukavsky,P.J., Otto,G.A., Lancaster,A.M., Sarnow,P., and Puglisi,J.D. (2000). Structures of two RNA domains essential for hepatitis C virus internal ribosome entry site function. Nat. Struct. Biol. *7*, 1105-1110.

Macdonald,M.A., Wodak,A.D., Dolan,K.A., van,B., I, Cunningham,P.H., and Kaldor,J.M. (2000). Hepatitis C virus antibody prevalence among injecting drug users at selected needle and syringe programs in Australia, 1995-1997. Collaboration of Australian NSPs. Med. J. Aust. *172*, 57-61.

Maggi, G., Armitano, S., Brambilla, L., Brenna, M., Cairo, M., Galvani, G., Gola, D., KomlaEbri, K., Marmondi, E., Perricone, G., Posca, M., Vegezzi, P. G., Vergani, C., and DeLeo, G. (1999). Hepatitis C infection in an Italian population not selected for risk factors. Liver 19[5], 427-431.

Mason,P.W., Bezborodova,S.V., and Henry,T.M. (2002). Identification and characterization of a cis-acting replication element (cre) adjacent to the internal ribosome entry site of foot-and-mouth disease virus. J. Virol. *76*, 9686-9694.

Mathews,D.H., Sabina,J., Zuker,M., and Turner,D.H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. *288*, 911-40.

McKnight, K. L. and Lemon, S. M. (1998). The rhinovirus type 14 genome contains an internally located RNA structure that is required for viral replication. RNA 4[12], 569-1584.

McOmish,F., Yap,P.L., Dow,B.C., Follett,E.A.C., Seed,C., Keller,A.J., Cobain,T.J., Krusius,T., Kolho,E., Naukkarinen,R., Lin,C., Lai,C., Leong,S., Medgyesi,G.A., Hejjas,M., Kiyokawa,H., Fukada,K., Cuypers,T., Saeed,A.A., Alrasheed,A.M., Lin,M., and Simmonds,P. (1994). Geographical distribution of hepatitis C virus genotypes in blood donors - an international collaborative survey. J. Clin. Microbiol. *32*, 884-892.

Mellits,K.H., Meredith,J.M., Rohll,J.B., Evans,D.J., and Almond,J.W. (1998). Binding of a cellular factor to the 3' untranslated region of the RNA genomes of entero- and rhinoviruses plays a role in virus replication. J. Gen. Virol. *79 (Pt 7)*, 1715-1723.

Melnick,J.L. (1982). Classification of hepatitis A virus as enterovirus type 72 and of hepatitis B virus as hepadnavirus type 1. Intervirology *18*, 105-106.

Mercier, B., Barclais, A., Botte, C., Cantalube, J. F., Coste, J., Defer, C., Gautreau, C., Giannoli, C., Halfon, P., Lepot, I., Loiseau, P., Martial, J., Montcharmont, P., Merel, P., Ouzan, D., Ravera, N., Follana, J., Cesaire, R., Janot, C., Lemaire, J. M., Demicco, P., Vezon, G., and Ferec, C. (1999). Prevalence of GBV C HGV RNA and GBV C HGV antibodies in French volunteer blood donors: Results of a collaborative study. Vox Sang 76[3], 166-169.

Miller,R.H. and Purcell,R.H. (1990). Hepatitis C virus shares amino acid sequence similarity with pestiviruses and flaviviruses as well as members of two plant virus supergroups. Proc. Natl. Acad. Sci. USA *87*, 2057-2061.

Minutello,M.A., Pileri,P., Unutmaz,D., Censini,S., Kuo,G., Houghton,M., Brunetto,M.R., Bonino,F., and Abrignani,S. (1993). Compartmentalization of T lymphocytes to the site of disease - intrahepatic CD4+ T-cells specific for the protein NS4 of hepatitis C virus in patients with chronic hepatitis C. J. Exp. Med. *178*, 17-25.

Missale, G., Bertoni, R., Lamonaca, V., Valli, A., Massari, M., Mori, C., Rumi, M. G., Houghton, M., Fiaccadori, F., and Ferrari, C. (1996). Different clinical behaviors of acute hepatitis C virus infection are associated with different vigor of the anti-viral cell-mediated immune response. J. Clin. Invest. 98[3], 706-714.

Mizushima,H., Hijikata,M., Asabe,S.I., Hirota,M., Kimura,K., and Shimotohno,K. (1994). Two hepatitis c virus glycoprotein e2 products with different c termini. J. Virol. *68*, 6215-6222.

Moaven,L.D., Locarnini,S.A., Bowden,D.S., Kim,J.P., Breschkin,A., McCaw,R., Yun,A., Wages,J., Jones,B., and Angus,P. (1997). Hepatitis G virus and fulminant hepatic failure: evidence for transfusion-related infection. J. Hepatol. *27*, 613-619.

Montelione,G.T., Zheng,D., Huang,Y.J., Gunsalus,K.C., and Szyperski,T. (2000). Protein NMR spectroscopy in structural genomics. Nat. Struct. Biol. *7 Suppl:982-5.*, 982-985.

Muerhoff,A.S., Leary,T.P., Simons,J.N., Pilot-Matias,T.J., Dawson,G.J., Erker,J.C., Chalmers,M.L., Schlauder,G.G., Desai,S.M., and Mushahwar,I.K. (1995). Genomic Organization of GB Viruses A and B: Two New Members of the Flaviviridae Associated with GB Agent Hepatitis. J. Virol. *69*, 5621-5630.

Murashima, S., Ide, T., Miyajima, I., Kumashiro, R., Ueno, T., Sakisaka, S., and Sata, M. (1999). Mutations in the NS5A gene predict response to interferon therapy in Japanese patients with chronic hepatitis C and cirrhosis. Scand J Infec Dis 31[1], 27-32.

Nakamura,T.M., Wang,Y.H., Zaug,A.J., Griffith,J.D., and Cech,T.R. (1995). Relative orientation of RNA helices in a group 1 ribozyme determined by helix extension electron microscopy. EMBO J. *14*, 4849-4859.

Nakao, H., Okamoto, H., Fukuda, M., Tsuda, F., Mitsui, T., Masuko, K., Lizuka, H., Miyakawa, Y., and Mayumi, M. (1997). Mutation rate of GB virus C hepatitis G virus over the entire genome and in subgenomic regions. Virology 233[1], 43-50.

Naoumov,N.V. (1999). Hepatitis C virus infection in Eastern Europe. J. Hepatol. *31 Suppl 1*, 84-87.

Nousbaum,J.B., Pol,S., Nalpas,B., Landais,P., Berthelot,P., Brechot,C., Gigou,M., Feray,C., Thiers,V., Okamoto,H., Mishiro,S., Poussin,K., Paterlini,P., Rumi,M., and Colombo,M. (1995). Hepatitis C virus type 1b (II) infection in France and Italy. Ann. Intern. Med. *122*, 161.

Nubling, C. M., Bialleck, H., Fursch, A. J., Scharrer, I., Schramm, W., Seifried, E., Schmidt, U., Staszewski, S., and Lower, J. (1997). Frequencies of GB virus C/hepatitis G virus genomes and of specific antibodies in German risk and non-risk populations. J. Med. Virol. 53[3], 218-224.

Ogata,N., Alter,H.J., Miller,R.H., and Purcell,R.H. (1991). Nucleotide sequence and mutation rate of the H strain of hepatitis C virus. Proc. Natl. Acad. Sci. USA *88*, 3392-3396.

Okamoto,H., Kojima,M., Okada,S.-I., Yoshizawa,H., Iizuka,H., Tanaka,T., Muchmore,E.E., Ito,Y., and Mishiro,S. (1992). Genetic drift of hepatitis C virus during an 8.2 year infection in a chimpanzee: variability and stability. Virology *190*, 894-899.

Okamoto, H., Nakao, H., Inoue, T., Fukuda, M., Kishimoto, J., Iizuka, H., Tsuda, F., Miyakawa, Y., and Mayumi, M. (1997). The entire nucleotide sequences of two GB virus C/hepatitis G virus isolates of distinct genotypes from Japan. J. Gen. Virol. 78, 737-745.

Okamoto,H., Okada,S., Sugiyama,Y., Kurai,K., Iizuka,H., Machida,A., Miyakawa,Y., and Mayumi,M. (1991). Nucleotide sequence of the genomic RNA of hepatitis C virus isolated from a human carrier: comparison with reported isolates for conserved and divergent regions. J. Gen. Virol. *72*, 2697-2704.

Olsthoorn,R.C. and Bol,J.F. (2001). Sequence comparison and secondary structure analysis of the 3' noncoding region of flavivirus genomes reveals multiple pseudoknots. RNA. *7*, 1370-1377.

Ostapowicz, G., Watson, K. J. R., Locarnini, S. A., and Desmond, P. V. (1998). Role of alcohol in the progression of liver disease caused by hepatitis C virus infection. Hepatology 27[6], 1730-1735.

Pessoa, M. G., Terrault, N. A., Detmer, J., Kolberg, J., Collins, M., Hassoba, H. M., and Wright, T. L. (1998). Quantitation of hepatitis G and C viruses in the liver: Evidence that hepatitis G virus is not hepatotropic. Hepatology 27[3], 877-880.

Peters,T., Schlayer,H.J., Preisler,S., Kopp,B., Berthold,H., Gerok,W., and Rasenack,J. (1993). Frequency of hepatitis C in acute post-transfusion hepatitis after open-heart surgery - a prospective study in 1,476 patients. J. Med. Virol. *39*, 139-145.

Pileri, P., Uematsu, Y., Campagnoli, S., Galli, G., Falugi, F., Petracca, R., Weiner, A. J., Houghton, M., Rosa, D., Grandi, G., and Abrignani, S. (1998). Binding of hepatitis C virus to CD81. Science 282[5390], 938-941.

Poovorawan, Y., Theamboonlers, A., Chongsrisawat, V., Seksarn, P., Jarvis, L., and Simmonds, P. (1998). High prevalence of hepatitis G virus infection in multiply transfused children with thalassaemia. J.Gastroenterol.Hepatol. 13[3], 253-256.

Prince,A.M., Brotman,B., Grady,G.F., Kuhns,W.J., Hazzi,C., Levine,R.W., and Millian,S.J. (1974). Long incubation post-transfusion hepatitis with evidence of exposure to hepatitis B virus. Lancet *ii*, 241-246.

Proutski,V., Gould,E.A., and Holmes,E.C. (1997). Secondary structure of the 3' untranslated region of flaviviruses: similarities and differences. Nucleic. Acids. Res. *25*, 1194-1202.

Ralston,R., Thudium,K., Berger,K., Kuo,C., Gervase,B., Hall,J., Selby,M., Kuo,G., Houghton,M., and Choo,Q.L. (1993). Characterization of hepatitis C virus envelope glycoprotein complexes expressed by recombinant vaccinia viruses. J. Virol. *67*, 6753-6761.

Rauscher,S., Flamm,C., Mandl,C.W., Heinz,F.X., and Stadler,P.F. (1997). Secondary structure of the 3'-noncoding region of flavivirus genomes: comparative analysis of base pairing probabilities. RNA. *3*, 779-791.

Ray,S.C., Arthur,R.R., Carella,A., Bukh,J., and Thomas,D.L. (2000). Genetic epidemiology of hepatitis C virus throughout egypt. J. Infect. Dis. *182*, 698-707.

Reyes,G.R., Purdy,M., Kim,J.S., Luk,K.C., Young,L.M., Fry,K.E., and Bradley,D.W. (1990). Isolation of cDNA from the virus responsible for enterically transmitted non-A, non-B hepatitis. Science *247*, 1335-1339.

Reynolds, J. E., Kaminski, A., Kettinen, H. J., Grace, K., Clarke, B. E., Carroll, A. R., Rowlands, D. J., and Jackson, R. J. Unique features of internal initiation of hepatitis C virus RNA translation. EMBO J. *14*, 6010-6020. 1995.

Rice,C.M., Lenches,E.M., Eddy,S.R., Shin,S.J., Sheets,R.L., and Strauss,J.H. (1985). Nucleotide sequence of yellow fever virus: implications for flavivirus gene expression and evolution. Science *229*, 726-733.

Rijnbrand,R., Bredenbeek,P., Vanderstraaten,T., Whetter,L., Inchauspe,G., Lemon,S., and Spaan,W. (1995). Almost the entire 5' non-translated region of hepatitis C virus is required for cap-independent translation. FEBS Lett. *365*, 115-119.

Rijnbrand,R., Bredenbeek,P.J., Haasnoot,P.C., Kieft,J.S., Spaan,W.J., and Lemon,S.M. (2001). The influence of downstream protein-coding sequence on internal ribosome entry on hepatitis C virus and other flavivirus RNAs. RNA. *7*, 585-597.

Rijnbrand, R. C. A., Abbink, T. E. M., Haasnoot, P. C. J., Spaan, W. J. M., and Bredenbeek, P. J. (1996). The influence of AUG codons in the hepatitis C virus 5' nontranslated region on translation and mapping of the translation initiation window. Virology 226[1], 47-56.

Rivas,E. and Eddy,S.R. (2000). Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. Bioinformatics. *16*, 583-605.

Robertus,J.D., Ladner,J.E., Finch,J.T., Rhodes,D., Brown,R.S., Clark,B.F., and Klug,A. (1974). Structure of yeast phenylalanine tRNA at 3 A resolution. Nature *250*, 546-551.

Roussel,J., Pillez,A., Montpellier,C., Duverlie,G., Cahour,A., Dubuisson,J., and Wychowski,C. (2003). Characterization of the expression of the hepatitis C virus F protein. J. Gen. Virol. *84*, 1751-1759.

Ruggieri, A., Argentini, C., Kouruma, F., Chionne, P., D'Ugo, E., Spada, E., Dettori, S., Sabbatani, S., and Rapicetta, M. (1996). Heterogeneity of hepatitis C virus genotype 2 variants in West Central Africa (Guinea Conakry). J. Gen. Virol. 77, 2073-2076.

Saeed,A.A., al Admawi,A.M., al Rasheed,A., Fairclough,D., Bacchus,R., Ring,C., and Garson,J. (1991). Hepatitis C virus infection in Egyptian volunteer blood donors in Riyadh. Lancet *338*, 459-460.

Saito, S., Tanaka, K., Kondo, M., Morita, K., Kitamura, T., Kiba, T., Numata, K., and Sekihara, H. (1997). Plus- and minus-stranded hepatitis G virus RNA in liver tissue and in peripheral blood mononuclear cells. Biochem. Biophys. Res. Com. 237[2], 288-291.

Sanchez,J.L., Sjogren,M.H., Callahan,J.D., Watts,D.M., Lucas,C., Abdel-Hamid,M., Constantine,N.T., Hyams,K.C., Hinostroza,S., Figueroa-Barrios,R., and Cuthie,J.C. (2000). Hepatitis C in Peru: risk factors for infection, potential iatrogenic transmission, and genotype distribution. Am. J. Trop. Med. Hyg. *63*, 242-248.

Santolini,E., Migliaccio,G., and Lamonica,N. (1994). Biosynthesis and biochemical properties of the hepatitis C virus core protein. J. Virol. *68*, 3631-3641.

Sarrazin, C., Herrmann, G., Roth, W. K., Lee, J. H., Marx, S., and Zeuzem, S. (1997). Prevalence and clinical and histological manifestation of hepatitis G/GBV-C infections in patients with elevated aminotransferases of unknown etiology. J. Hep. 27[2], 276-283.

Sathar,M.A., Soni,P.N., Pegoraro,R., Simmonds,P., Smith,D.B., Dhillon,A.P., and Dusheiko,G.M. (1999). A new variant of GB virus C/hepatitis G virus (GBV-C/HGV) from South Africa. Virus. Res. *64*, 151-160.

Sbardellati, A., Scarselli, E., Tomei, L., Kekule, A. S., and Traboni, C. (1999). Identification of a novel sequence at the 3 ' end of the GB virus B genome. J Virol 73[12], 10546-10550.

Scallan, M. F., Clutterbuck, D., Jarvis, L. M., Scott, G. R., and Simmonds, P. (1998). Sexual transmission of GB virus-C/hepatitis G virus. J.Med Virol. 55, 203-208.

Schreiber, G. B., Busch, M. P., Kleinman, S. H., and Korelitz, J. J. (1996). The risk of transfusion-transmitted viral infections. N. Engl. J. Med. 334[26], 1685-1690.

Seeff, L. B. (1998). The natural history of hepatitis C - A quandary. Hepatology 28[6], 1710-1712.

Seffens,W. and Digby,D. (1999). mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. Nucleic Acids Res. *27*, 1578-1584.

SenGupta,D.J., Zhang,B., Kraemer,B., Pochart,P., Fields,S., and Wickens,M. (1996). A three-hybrid system to detect RNA-protein interactions in vivo. Proc. Natl. Acad. Sci. U. S. A. *93*, 8496-501.

Shi,P.Y., Li,W., and Brinton,M.A. (1996). Cell proteins bind specifically to West Nile virus minus-strand 3' stem-loop RNA. J. Virol. *70*, 6278-6287.

Shimizu, Y. K., Igarashi, H., Kiyohara, T., Cabezon, T., Farci, P., Purcell, R. H., and Yoshikura, H. (1996). A hyperimmune serum against a synthetic peptide corresponding to the hypervariable region 1 of hepatitis C virus can prevent viral infection in cell cultures. Virology 223[2], 409-412.

Shimizu,Y.K., Weiner,A.J., Rosenblatt,J., Wong,D.C., Shapiro,M., Popkin,T., Houghton,M., Alter,H.J., and Purcell,R.H. (1990). Early events in hepatitis C virus infection of chimpanzees. Proc. Natl. Acad. Sci. USA *87*, 6441-6444.

Shimoike, T., Mimori, S., Tani, H., Matsuura, Y., and Miyamura, T. (1999). Interaction of hepatitis C virus core protein with viral sense RNA and suppression of its translation. J Virol 73[12], 9718-9725.

Silini,E., Bono,F., Cerino,A., Piazza,V., Solcia,E., and Mondelli,M.U. (1993). Virological features of hepatitis C virus infection in hemodialysis patients. J. Clin. Microbiol. *31*, 2913-2917.

Simmonds,P., Alberti,A., Alter,H.J., Bonino,F., Bradley,D.W., Brechot,C., Brouwer,J.T., Chan,S.W., Chayama,K., Chen,D.S., Choo,Q.L., Colombo,M., Cuypers,H.T.M., Date,T., Dusheiko,G.M., Esteban,J.I., Fay,O., Hadziyannis,S.J., Han,J., Hatzakis,A., Holmes,E.C., Hotta,H., Houghton,M., Irvine,B., Kohara,M., Kolberg,J.A., Kuo,G., Lau,J.Y.N., Lelie,P.N., Maertens,G., McOmish,F., Miyamura,T., Mizokami,M., Nomoto,A., Prince,A.M., Reesink,H.W., Rice,C.,

Roggendorf,M., Schalm,S.W., Shikata,T., Shimotohno,K., Stuyver,L., Trepo,C., Weiner,A., Yap,P.L., and Urdea,M.S. (1994a). A proposed system for the nomenclature of hepatitis C viral genotypes. Hepatology *19*, 1321-1324.

Simmonds,P., McOmish,F., Yap,P.L., Chan,S.W., Lin,C.K., Dusheiko,G., Saeed,A.A., and Holmes,E.C. (1993). Sequence variability in the 5' non coding region of hepatitis C virus: identification of a new virus type and restrictions on sequence diversity. J. Gen. Virol. *74*, 661-668.

Simmonds, P., Mellor, J., Sakuldamrongpanich, T., Nuchaprayoon, C., Tanprasert, S., Holmes, E. C., and Smith, D. B. (1996). Evolutionary analysis of variants of hepatitis C virus found in South-East Asia: Comparison with classifications based upon sequence similarity. J. Gen. Virol. 77, 3013-3024.

Simmonds,P. and Smith,D.B. (1999a). Hepatitis C and G viruses - Old or New? In HIV and the new viruses, A.G.Dalgleish and R.A.Weiss, eds. (San Diego: Academic Press), pp. 459-480.

Simmonds,P. and Smith,D.B. (1999b). Structural constraints on RNA virus evolution. J. Virol. *73*, 5787-5794.

Simmonds,P., Smith,D.B., McOmish,F., Yap,P.L., Kolberg,J., Urdea,M.S., and Holmes,E.C. (1994b). Identification of genotypes of hepatitis C virus by sequence comparisons in the core, E1 and NS-5 regions. J. Gen. Virol. *75*, 1053-1061.

Simons,J.N., Desai,S.M., Schultz,D.E., Lemon,S.M., and Mushahwar,I.K. (1996). Translation initation in GB viruses A and C: evidence for internal ribosome entry and implications for genome organisation. J. Virol. *70*, 6126-6135.

Simons,J.N., Leary,T.P., Dawson,G.J., Pilot-Matias,T.J., Muerhoff,A.S., Schlauder,G.G., Desai,S.M., and Mushahwar,I.K. (1995a). Isolation of novel virus-like sequences associated with human hepatitis. Nature Med. *1*, 564-569.

Simons,J.N., Pilot-Matias,T.J., Leary,T.P., Dawson,G.J., Desai,S.M., Schlauder,G.G., Muerhoff,A.S., Erker,J.C., Buijk,S.L., Chalmers,M.L., Vansant,C.L., and Mushahwar,I.K. (1995b). Identification of two flavivirus-like genomes in the GB hepatitis agent. Proc. Natl. Acad. Sci. USA *92*, 3401-3405.

Smith,D.B., Basaras,M., Frost,S., Haydon,D., Cuceanu,N., Prescott,L., Kamenka,C., Millband,D., Sathar,M.A., and Simmonds,P. (2000). Phylogenetic analysis of GBV-C/hepatitis G virus. J. Gen. Virol. *81*, 769-780.

Smith, D. B., Cuceanu, N., Davidson, F., Jarvis, L. M., Mokili, J. L. K., Hamid, S., Ludlam, C. A., and Simmonds, P. (1997b). Discrimination of hepatitis G virus/GBV-C geographical variants by analysis of the 5' non-coding region. J. Gen. Virol. 78, 1533-1542.

Smith,D.B., Mellor,J., Jarvis,L.M., Davidson,F., Kolberg,J., Urdea,M., Yap,P.L., Simmonds,P., Conradie,J.D., Neill,A.G.S., Dusheiko,G.M., Kew,M.C., Crookes,R., Koshy,A., Lin,C.K., Lai,C., Murraylyon,I.M., Elguneid,A., Gunaid,A.A., Yemen,T., Yemen,S., Mutimer,D., Ahmed,M., Nuchprayoon,C., Tanprasert,S., Preston,F.E.,

Makris,M., Chuansumrit,A., Mahasandana,C., Pritchard,D., Riley,E., Greenwood,B.M., Saeed,A.A., Alrasheed,A.M., Saleh,M.G., Mcfarlane,I., Tibbs,C., Williams,R., Power,J., Lawlor,E., and Kiyokawa,H. (1995). Variation of the hepatitis C virus 5' non-coding region: implications for secondary structure, virus detection and typing. J. Gen. Virol. *76*, 1749-1761.

Smith, D. B., Pathirana, S., Davidson, F., Lawlor, E., Power, J., Yap, P. L., and Simmonds, P. (1997b). The origin of hepatitis C virus genotypes. J. Gen. Virol. 78, 321-328. 1997.

Smith,D.B. and Simmonds,P. (1997a). Characteristics of nucleotide substitution in the hepatitis C virus genome: constraints on sequence change in coding regions at both ends of the genome. J. Mol. Evol. *45*, 238-246.

Smith,R.M., Walton,C.M., Wu,C.H., and Wu,G.Y. (2002). Secondary structure and hybridization accessibility of hepatitis C virus 3'-terminal sequences. J. Virol. *76*, 9563-9574.

Stamenkovic,G., Zerjav,S., Velickovic,Z.M., Krtolica,K., Samardzija,V.L., Jemuovic,L., Nozic,D., and Dimitrijevic,B. (2000). Distribution of HCV genotypes among risk groups in Serbia. Eur. J. Epidemiol. *16*, 949-954.

Tacke, M., Schmolke, S., Schlueter, V., Sauleda, S., Esteban, J. I., Tanaka, E., Kiyosawa, K., Alter, H. J., Schmitt, U., Hess, G., OfenlochHaehnle, B., and Engel, A. M. (1997). Humoral immune response to the E2 protein of hepatitis G virus is associated with long-term recovery from infection and reveals a high frequency of hepatitis G virus exposure among healthy blood donors. Hepatology 26[6], 1626-1633.

Takahashi,K., Kishimoto,S., Yoshizawa,H., Okamoto,H., Yoshikawa,A., and Mishiro,S. (1992). p26-protein and 33-nm particle associated with nucleocapsid of hepatitis-c virus recovered from the circulation of infected hosts. Virology *191*, 431-434.

Takamizawa,A., Mori,C., Fuke,I., Manabe,S., Murakami,S., Fujita,J., Onishi,E., Andoh,T., Yoshida,I., and Okayama,H. (1991). Structure and organization of the hepatitis C virus genome isolated from human carriers. J. Virol. *65*, 1105-1113.

Tanaka,T., Kato,N., Cho,M.J., and Shimotohno,K. (1995). A novel sequence found at the end of the 3' terminus of hepatitis C virus genome. Biochem. Biophys. Res. Commun. *215*, 744-749.

Tanaka, T., Kato, N., Cho, M. J., Sugiyama, K., and Shimotohno, K. (1996). Structure of the 3' terminus of the hepatitis C virus genome. J. Virol. 70[5], 3307-3312.

Tanaka, Y., Mizokami, M., Orito, E., Ohba, K., Kato, T., Kondo, Y., Mboudjeka, I., Zekeng, L., Kaptue, L., Bikandou, B., MPele, P., Takehisa, J., Hayami, M., Suzuki, Y., and Gojobori, T. (1998a). African origin of GB virus C hepatitis G virus. FEBS Lett. 423[2], 143-148.

Tanaka, Y., Mizokami, M., Orito, E., Ohba, K. I., Nakano, T., Kato, T., Kondo, Y., Ding, X., Ueda, R., Sonoda, S., Tajima, K., Miura, T., and Hayami, M. (1998b). GB virus C/hepatitis G virus infection among Colombian native Indians. Am.J.Trop.Med.Hyg. 59[3], 462-467.

Tanji,Y., Hijikata,M., Satoh,S., Kaneko,T., and Shimotohno,K. (1995). Hepatitis C virus-encoded nonstructural protein NS4a has versatile functions in viral protein processing. J. Virol. *69*, 1575-1581.

Tassopoulos,N.C., Hatzakis,A., Delladetsima,I., Koutelou,M.G., Todoulos,A., and Miriagou,V. (1992). Role of hepatitis C virus in acute non-A, non-B hepatitis in Greece: A 5-year prospective study. Gastroenterology *102*, 969-972.

Tassopoulos,N.C., Krawczynski,K., Hatzakis,A., Katsoulidou,A., Delladetsima,I., Koutelou,M.G., and Trichopoulos,D. (1994). Role of hepatitis e virus in the etiology of community- acquired non-A, non-B hepatitis in Greece - case report. J. Med. Virol. *42*, 124-128.

Taylor,A., Goldberg,D., Hutchison,S., Cameron,S., Gore,S., McMenamin,J., Green,S., Pithie,A., and Fox,R. (2000). Prevalence of hepatitis C virus infection amongst injecting drug users in Glasgow 1990-1996: are current harm reduction strategies working? J. Inf. *40*, 176-183.

Thomas, D. L., Vlahov, D., Alter, H. J., Hunt, J. C., Marshall, R., Astemborski, J., and Nelson, K. E. (1998). Association of antibody to GB virus C (hepatitis G virus) with viral clearance and protection from reinfection. J. Inf. Dis. 177[3], 539-542.

Thomssen,R., Bonk,S., and Thiele,A. (1993). Density heterogeneities of hepatitis C virus in human sera due to the binding of beta-lipoproteins and immunoglobulins. Med. Microbiol. Immunol. (Berl) *182*, 329-334.

Tisminetzky,S.G., Gerotto,M., Pontisso,P., Chemello,L., Ruvoletto,M.G., Baralle,F., and Alberti,A. (1994). Genotypes of hepatitis C virus in Italian patients with chronic hepatitis C. Int. Hepatol. Commun. *2*, 105-112.

Trepo,C. and Pradat,P. (1999). Hepatitis C virus infection in Western Europe. J. Hepatol. *31 Suppl 1*, 80-83.

Tsukiyama Kohara,K., Iizuka,N., Kohara,M., and Nomoto,A. (1992). Internal ribosome entry site within hepatitis C virus RNA. J. Virol. *66*, 1476-1483.

Tsukuma,H., Hiyama,T., Tanaka,S., Nakao,M., Yabuuchi,T., Kitamura,T., Nakanishi,K., Fujimoto,I., Inoue,A., Yamazaki,H., and Kawashima,T. (1993). Risk factors for hepatocellular carcinoma among patients with chronic liver disease. N. Engl. J. Med. *328*, 1797-1801.

Tucker, T. J., Smuts, H., Eickhaus, P., Robson, S. C., and Kirsch, R. E. (1999). Molecular characterization of the 5 ' non-coding region of South African GBV-C/HGV isolates: Major deletion and evidence for a fourth genotype. J Med Virol 59[1], 52-59.

Tucker,T.J., Smuts,H.E., Eedes,C., Knobel,G.D., Eickhaus,P., Robson,S.C., and Kirsch,R.E. (2000). Evidence that the GBV-C/hepatitis G virus is primarily a lymphotropic virus. J. Med. Virol. *61*, 52-58.

Tuplin,A., Wood,J., Evans,D.J., Patel,A.H., and Simmonds,P. (2002). Thermodynamic and phylogenetic prediction of RNA secondary structures in the coding region of hepatitis C virus. RNA *8*, 824-841.

Viazov, S., Riffelmann, M., Khoudyakov, Y., Fields, H., Varenholz, C., and Roggendorf, M. (1997). Genetic heterogeneity of hepatitis G virus isolates from different parts of the world. J. Gen. Virol. 78, 577-581.

Walter,F., Murchie,A.I., Thomson,J.B., and Lilley,D.M. (1998). Structure and activity of the hairpin ribozyme in its natural junction conformation: effect of metal ions. Biochem. *37*, 14195-14203.

Wang,C.Y., Sarnow,P., and Siddiqui,A. (1993). Translation of human hepatitis C virus RNA in cultured cells is mediated by an internal ribosome-binding mechanism. J. Virol. *67*, 3338-3344.

Wang, H. and Eckels, D. D. (1999). Mutations in immunodominant T cell epitopes derived from the nonstructural 3 protein of hepatitis C virus have the potential for generating escape variants that may have important consequences for T cell recognition. J. Immunol. 162[7], 4177-4183.

Wang H, Rijnbrand RC, Lemon SM. (2000). Core protein-coding sequence, but not core protein, modulates the efficiency of cap-independent translation directed by the internal ribosome entry site of hepatitis C virus. J Virol. *74*, 11347-58.

Wang,Y.H., Murphy,F.L., Cech,T.R., and Griffith,J.D. (1994). Visualization of a tertiary structural domain of the Tetrahymena group I intron by electron microscopy. J. Mol. Biol. *236*, 64-71.

Weiner,A.J., Brauer,M.J., Rosenblatt,J., Richman,K.H., Tung,J., Crawford,K., (Bonino,F., Saracco,G., Choo,Q.L., Houghton,M., and Han,J.H. (1991). Variable and hypervariable domains are found in the regions of HCV corresponding to the flavivirus envelope and NS1 proteins and the pestivirus envelope glycoproteins. Virology *180*, 842-848.

Westin, J., Lindh, M., Lagging, L. M., Norkrans, G., and Wejstal, R. (1999). Chronic hepatitis C in Sweden: Genotype distribution over time in different epidemiological settings. Scand J Infec Dis 31[4], 355-358.

Witwer,C., Rauscher,S., Hofacker,I.L., and Stadler,P.F. (2001). Conserved RNA secondary structures in Picornaviridae genomes. Nucleic Acids Res. *29*, 5079-5089.

Wolk,B., Sansonno,D., Krausslich,H.G., Dammacco,F., Rice,C.M., Blum,H.E., and Moradpour,D. (2000). Subcellular localization, stability, and trans-cleavage competence of the hepatitis C virus NS3-NS4A complex expressed in tetracycline-regulated cell lines. J. Virol. *74*, 2293-2304.

Wood,J., Frederickson,R.M., Fields,S., and Patel,A.H. (2001). Hepatitis C virus 3'X region interacts with human ribosomal proteins. J. Virol. *75*, 1348-58.

Workman,C. and Krogh,A. (1999). No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. Nucleic. Acids. Res. *27*, 4816-22.

Xiang,J., Wunschmann,S., Schmidt,W., Shao,J., and Stapleton,J.T. (2000). Full-length GB virus C (Hepatitis G virus) RNA transcripts are infectious in primary CD4-positive T cells. J. Virol. *74*, 9125-9133.

Xiang, J. H., Klinzman, D., McLinden, J., Schmidt, W. N., Labrecque, D. R., Gish, R., and Stapleton, J. T. (1998). Characterization of hepatitis G virus (GB-C virus) particles: Evidence for a nucleocapsid and expression of sequences upstream of the E1 protein. J. Virol. 72[4], 2738-2744.

Xu,L.Z., Larzul,D., Delaporte,E., Brechot,C., and Kremsdorf,D. (1994). Hepatitis c virus genotype 4 is highly prevalent in central africa (gabon). J. Gen. Virol. *75*, 2393-2398.

Yanagi, M., StClaire, M., Emerson, S. U., Purcell, R. H., and Bukh, J. (1999). In vivo analysis of the 3 ' untranslated region of the hepatitis C virus after in vitro mutagenesis of an infectious cDNA clone. Proc Nat Acad Sci USA 96[5], 2291-2295.

Yasui, K., Wakita, T., Tsukiyamakohara, K., Funahashi, S., Ichikawa, M., Kajita, T., Moradpour, D., Wands, J. R., and Kohara, M. (1998). The native form and maturation process of hepatitis C virus core protein. J. Virol. 72[7], 6048-6055.

Yoshiba,M., Dehara,K., Inoue,K., Okamoto,H., and Mayumi,M. (1994). Contribution of hepatitis c virus to non-a, non-b fulminant hepatitis in japan. Hepatology *19*, 829-835.

Yoshiba,M., Okamoto,H., and Mishiro,S. (1995). Detection of the GBV-C hepatitis virus genome in serum from patients with fulminant hepatitis of unkown aetiology. Lancet *346*, 1131-1132.

Yu, H. Y., Grassmann, C. W., and Behrens, S. E. (1999). Sequence and structural elements at the 3 ' terminus of bovine viral diarrhea virus genomic RNA: Functional role during RNA replication. J Virol 73[5], 3638-3648.

Yu,M.L., Chuang,W.L., Chen,S.C., Dai,C.Y., Hou,C., Wang,J.H., Lu,S.N., Huang,J.F., Lin,Z.Y., Hsieh,M.Y., Tsai,J.F., Wang,L.Y., and Chang,W.Y. (2001). Changing prevalence of hepatitis C virus genotypes: molecular epidemiology and clinical implications in the hepatitis C virus hyperendemic areas and a tertiary referral center in Taiwan. J. Med. Virol. *65*, 58-65.

Yunoki,M., Tsujikawa,M., Urayama,T., Sasaki,Y., Morita,M., Tanaka,H., Hattori,S., Takechi,K., and Ikuta,K. (2003). Heat sensitivity of human parvovirus B19. Vox Sang. *84*, 164-169.

Zanetti, A. R., Tanzi, E., Romano, L., Principi, N., Zuin, G., Minola, E., Zapparoli, B., Palmieri, M., Marini, A., Ghisotti, D., Friedman, P., Hunt, J., and Laffler, T. (1998). Multicenter trial on mother-to-infant transmission of GBV-C virus. J. Med. Virol. 54[2], 107-112.

Zhang, X. H., Shinzawa, H., Shao, L., Ishibashi, M., Jiang, Q. H., Saito, K., Misawa, H., Togashi, H., and Takahashi, T. (1998). Epidemiological study and genetic analysis of GB virus C infection in general population from an area endemic for hepatitis C. J. Med. Virol. 54[4], 237-242.

Zuckerman,A.J. (1996). Alphabet of hepatitis viruses. Lancet *347*, 558-559.

Zuker,M. (2000). Calculating nucleic acid secondary structure. Curr. Opin. Struct. Biol. *10*, 303-310.

Zuker,M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. *31*, 3406-3415.

APPENDIX

2.1 POLYMERASE CHAIN REACTION

2.1.1 EXTRACTION OF VIRAL RNA FROM SERUM SAMPLES

Lysis buffer: 50mM Tris-Hcl (ph 8.0)

100mM NaCl

1mM EDTA

0.5% sodium-n-lauroylsarcosine

1 mg/ml Proteinase K

40 µg/ml polyadenylic acid

Phenol (Rathburn Chemical Ltd)

Chloroform

Isoamyalcohol (BDH)

Sodium acetate (pH 5.2)

100% ethanol

2.1.2 REVERSE TRANSCRIPTION OF VIRAL RNA

RT buffer: 50 mM Tris-HCl pH 8.3

50 mM KCl

10 mM $MgCl_2$

10 mM DTT

5 mM spermidine

Nucleotide triphosphate mix (100 mM stock solution; diluted to 4 mM; Promega)

RNasin (Ribonuclease inhibitor 20 U/µl; Promega)

Avian Myeoblastosis Virus Reverse Transcriptase (AMV; Promega; 1 U/µl)

2.1.3 and 2.2.5 PCR AMPLIFICATION

10 × PCR reaction buffer: 10 mM Tris-HCl (pH 9.0)

50 mM KCl

1.5 mM $MgCl_2$

0.1% Triton X-100

Nucleotide triphosphate mix (100 mM stock solution; diluted to 3 mM; Promega)

Taq polymerase (5 U/µl; Promega)

2.1.4 ANALYSIS OF PCR PRODUCT

1% agarose gel: 1.5g agarose (Flowgen)

150 ml 1 × TAE

0.07 µg/ml ethidium bromide (Sigma)

50 × TAE: 242g Tris base (Sigma)

57.1 ml glacial acetic acid (BDH)

37.2g EDTA (Sigma)

Make up to 1L with distilled water

1 × TBE used as electrophoresis buffer in gel tank

2.2 CLONING OF PCR PRODUCT

Luria Broth (LB) medium: 10 g Bacto-tryptone (GIBCO-BRL)

5 g Bacto-yeast extract (GIBCO-BRL)

10 g NaCl

Make up to 1L with distilled water sterilize by autoclaving

Make up to 1L with distilled water: Add 15 g of agar (GIBCO-BRL) to 1 litre of water and sterilise by autoclaving. Allow to cool before adding 0.5 nM IPTG: 80μg/ml X-Gal and 100 μg/ml ampicillin.

2.2.7 DNA SEQUENCING

Termination mix: 7.5 μM each of dATP, dCTP dGTP and dTTP

0.5 μl [α-33P] ddATP, ddCTP ddGTP or ddTTP (0.45 μCi/μl) (Amersham)

Reaction mixture: 260 mM Tris-HCl, pH 9.5

65 mM $MgCl_2$

0.15 mM antisense primer

4 U of Thermo Sequenase polymerase 4 U/µl (USB)

Denaturing stop solution: 95 % formamide

20 mM EDTA

0.05 bromophenol blue

0.05% xylene cyanol

5% denaturing acrylamide gel: 21 g urea (Ana-BDH)

5% v/v Ultrapure Sequagel concentrate (acryl:bis-acryl = 19:1) (National Diagnostics)

5 ml of 10× Sanger TBE

0.05 g ammonium persulphate

20 µl of TEMED (N,N,N,Ntetramethylethlediamine) (Sigma)

Make up to 20 µl with distilled water

10 × TBE: 324g Tris base

85 g Boric acid

19 g EDTA

Make up to 2L with distilled water

2.6. THERMODYNAMIC PREDICTION OF FOLDING FREE ENERGY (FFE) AND STRUCTURAL PREDICTION

Sequences analysed for comparative structural predictions (chapter 4): Sequences analysed for FFED are shown underlined (chapter 3).

HCV (genotype 1a): AF011751, AF064490, AF511948, AF511950, AY051292, D50409, HEC278830, HPCCGS, HPCHCJ1, HPCPLYPRE,

HCV (genotype 1b): AB049087, AB049088, AB049089, AB049090, AB049091, AB049092, AB049093, AB049094, AB049095, AB049096, AB049097, AB049098, AB049099, AB049100, AB049101, AB080299, AF139594, AF165045, AF165047, AF165049, AF165051, AF165053, AF165055, AF165057, AF165059, AF165061, AF165063, AF207752, AF207753, AF207754, AF207756, AF207757, AF207758, AF207759, AF207760, AF207761, AF207762, AF207763, AF207764, AF207765, AF207766, AF207767, AF207768, AF207769, AF207770, AF207771, AF207772, AF207773, AF207774, AF208024, AF313916, AF333324, AF356827, AF483269, AF511949, AY045702, D85516, D89815, D89872, HCJ238799, HCU01214, HCU45476, HCV132996, HCVJK1G, HCVPOLYP, HPCCGENOM, HPCGENANTI, HPCHUMR, HPCJ491, HPCJCG, HPCJRNA, HPCJTA, HPCK1R1, HPCK1R2, HPCK1R3, HPCPP, HPCRNA, HPCUNKCDS, HPVHCVN,

HCV (genotype 2a): AB030907, AB031663, AB047639, AB047640, AB047641, AB047642, AB047643, AB047644, AB047645, AF169002, AF169003, AF169004, AF169005, AF177036, AF238481, AF238482, AF238483, AF238484, AF238485, AF238486, HPCJ8G, HPCPOLP,

HCV (genotype 3a): AF046866, HCVCENS1, HPCEGS, HPCFG, HPCJK046E2, HPCK3A,

HCV (genotype 4a): HCV4APOLY,

HCV (genotype 5a): HCV1480,

HCV (genotype 6a): D84262, D84263, D84264, D84265, HCV12083, HPCJK049E1

HGV/GBV-C (genotype 1): U36380, AB013500, AB003291

HGV/GBV-C (genotype 2): AB013501, AF031829, D90600, D87255, AF104403, U63715, U44402, AB003289, U45966

HGV/GBV-C (genotype 3): D90601, D87263, AB008342, AB003288, AB003293, AF006500, AB003290, U94695

HGV/GBV-C (genotype 4): AB018667, AB021287, AB003292,

HGV/GBV-C_{CPZ}: AF070476

GBV-A: NC_001837, U22303, AF023425, AF023424

GBV-B: NC-001655

The coding regions of the following mammalian sequences were used as FFED negative controls (chapter 3):

Actin: BC015695

Albumin: AF116645, X84842, Y17729, AB006197, XY4045, M90463, U01222;

Alphaglobin: V00493, XO5289, M12158, X02008, M17083.

HLA DRw12 beta 1-chain

The following sequences were used as FFED positive controls (chapter 3):

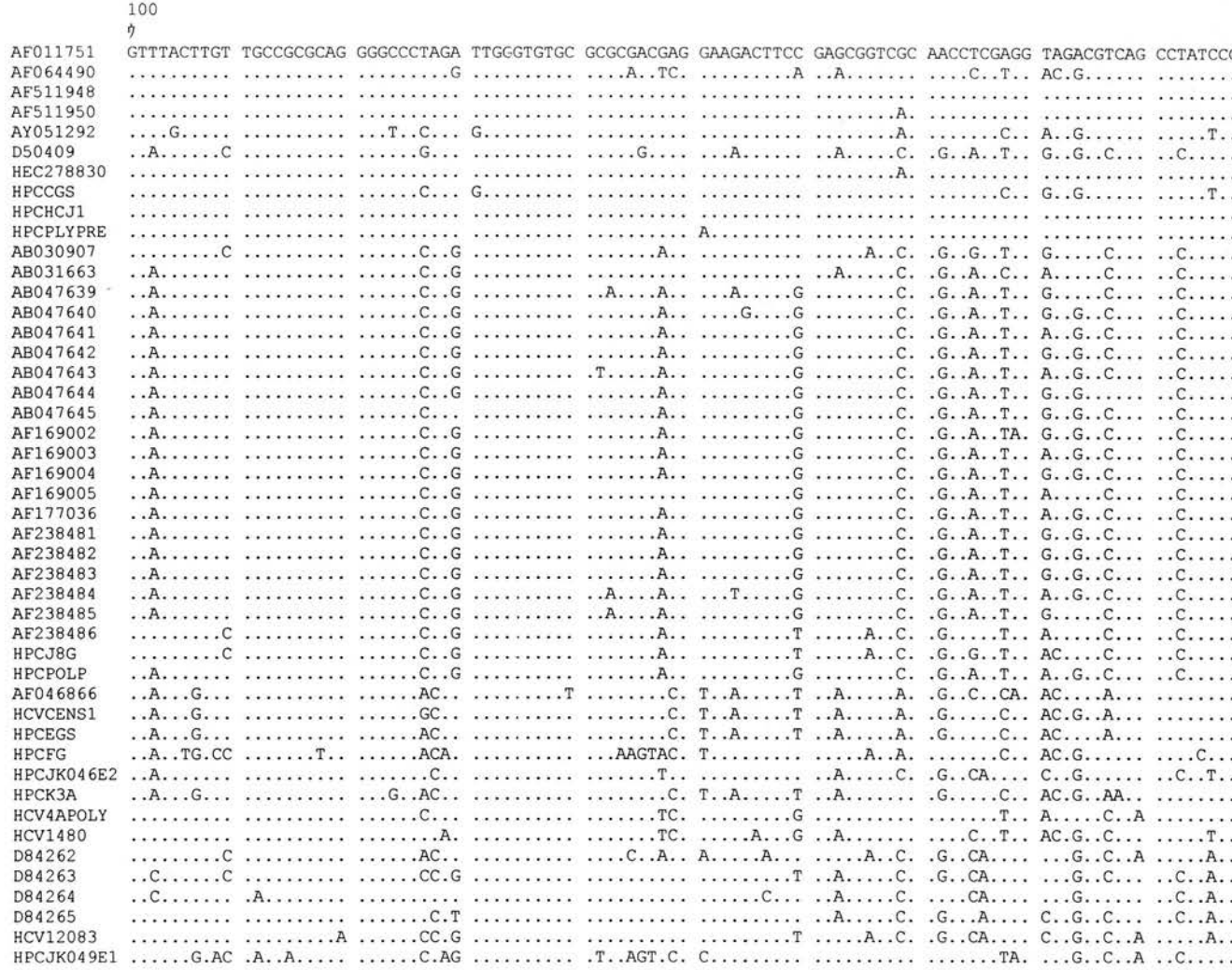
Plant viroid: X53715, U23058, AJ247123, L78463, X76846

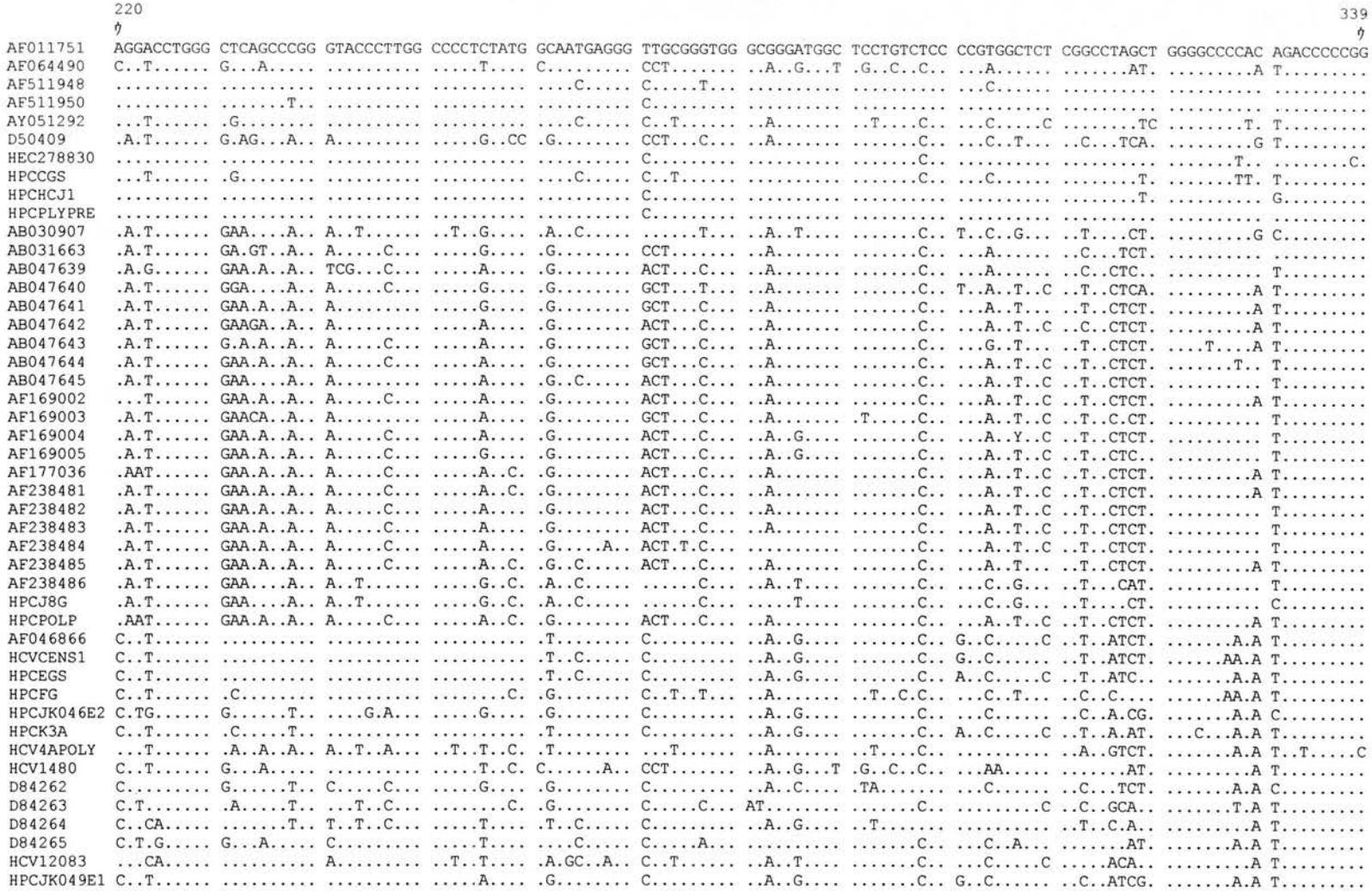
Delta virus: AF098261, AJ000558, M31012, D01075, M28267

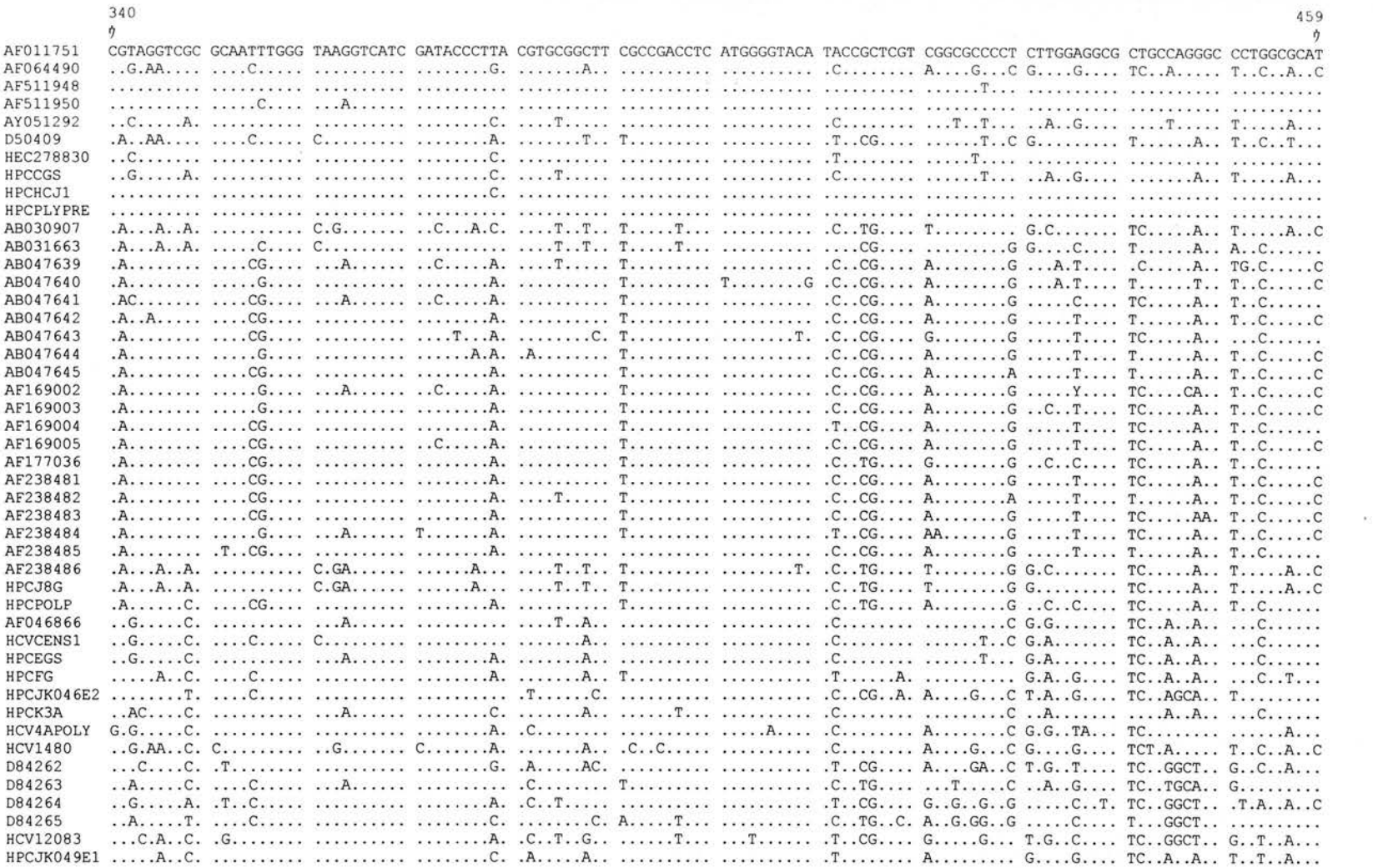
Sequences alignments created for comparative structural predictions (chapter 4) (genotype 1b is not shown in the HCV sequence alignment):

HCV: Core gene

```
           -20                                                
```


100
219
AF011751 GTTTACTTGT TGCCGCGCAG GGGCCCTAGA TTGGGTGTGC GCGCGACGAG GAAGACTTCC GAGCGGTCGC AACCTCGAGG TAGACGTCAG CCTATCCCCA AGGCACGTCG GCCCGAGGGC
AF064490 ....G ....A..TC. .....A ..A..... ....C..T.. AC.G..... ....G..C.A ....AC....
AF511948 ....G..... .....
AF511950 ....A. ....
AY051292 ....G..... ....T..C... G..... ....A. .....C.. A..G..... .....T.... ....C..C.. A.....
D50409 ..A.....C ....G... ....G.... ...A..... ..A.....C. .G..A..T.. G..G..C... ..C.....T. .A.AT..G.. CT..ACT...
HEC278830 ....A. ....G..... .....
HPCCGS ....C... G..... ....C.. G..G..... .....T.... ....C..C.. A.....A
HPCHCJ1 ....TG..... .....
HPCPLYPRE .... A..... ....T..... .....
AB030907 .....C ....C..G ....A.. ....A..C. .G..G..T.. G.....C... ..C.....G. .A.AT..G.. CT..ACC...
AB031663 ..A..... ....C..G ....A.....C. .G..A..C.. A.....C... ..C..... .A.AT..G.. TT...CT...
AB047639 ..A..... ....C..G ....A....A.. ...A.....G .....C. .G..A..T.. G.....C... ..C..... .A.AT..G.. CT..ACT...
AB047640 ..A..... ....C..G ....A.. ....G....G .....C. .G..A..T.. G..G..C... ..C..... .ACAT..G.. CT..ACT...
AB047641 ..A..... ....C..G ....A.. ....G. .....C. .G..A..T.. A..G..C... ..C.....T. .A.AT..G.. CT..ACT...
AB047642 ..A..... ....C..G ....A.. ....G. .....C. .G..A..T.. G..G..C... ..C..... .A.AT..G.. CT..ACT...
AB047643 ..A..... ....C..G ..... .T.....A.. ....G. .....C. .G..A..T.. A..G..C... ..C..... .A.AT..G.. CT..ACT...
AB047644 ..A..... ....C..G ....A.. ....G. .....C. .G..A..T.. G..G..... ..C.....T. .A.AT..G.. CT..ACT...
AB047645 ..A..... ....C... ....A.. ....G. .....C. .G..A..T.. G..G..C... ..C.....T. .A.AT..G.. CT..ACT...
AF169002 ..A..... ....C..G ....A.. ....G. .....C. .G..A..TA. G..G..C... ..C..... .A.AT..G.. CT..ACT...
AF169003 ..A..... ....C..G ....A.. ....G. .....C. .G..A..T.. A..G..C... ..C.....T. .ACAT..G.. CT..ACT...
AF169004 ..A..... ....C..G ....A.. ....G. .....C. .G..A..T.. G..G..C... ..C..... .A.AT..G.. CT..ACT...
AF169005 ..A..... ....C..G ....G. .....C. .G..A..T.. A.....C... ..C..... .A.AT..G.. CT..ACT...
AF177036 ..A..... ....C..G ....A.. ....G. .....C. .G..A..T.. A..G..C... ..C.....T. .A.AT..G.. CT..ACT...
AF238481 ..A..... ....C..G ....A.. ....G. .....C. .G..A..T.. G..G..C... ..C.....T. .A.AT..G.. CT..ACT...
AF238482 ..A..... ....C..G ....A.. ....G. .....C. .G..A..T.. G..G..C... ..C..... .A.AT..G.. CT..ACT...
AF238483 ..A..... ....C..G ....A.. ....G. .....C. .G..A..T.. G..G..C... ..C..... .A.AT..G.. CT..ACT...
AF238484 ..A..... ....C..G ....A....A.. ...T..... .....G. .....C. .G..A..T.. A..G..C... ..C.....T. .A.AT..G.. CT..ACT...
AF238485 ..A..... ....C..G ....A....A.. ....G. .....C. .G..A..T.. G.....C... ..C..... .A.AT..G.. CT..ACT...
AF238486 .....C ....C..G ....A.. ....T ....A..C. .G.....T.. A.....C... ..C.....G. .A.AT..G.. CT..ACC...
HPCJ8G .....C ....C..G ....A.. ....T ....A..C. .G..G..T.. AC....C... ..C.....G. .A.AT..G.. CT..ACC...
HPCPOLP ..A..... ....C..G ....A.. ....G ....C. .G..A..T.. A..G..C... ..C.....T. ...AT..G.. CT..ACT...
AF046866 ..A...G... ....AC.. .....T ....C. T..A.....T ..A.....A. .G..C..CA. AC....A... ....G....A .AG..GA...
HCVCENS1 ..A...G... ....GC.. ....C. T..A.....T ..A.....A. .G.....C.. AC.G..A... ....G..... .AG...A...
HPCEGS ..A...G... ....AC.. ....C. T..A.....T ..A.....A. .G.....C.. AC....A... ....G..... .AG...A...
HPCFG ..A..TG.CC .....T.. .....ACA. ....AAGTAC. T..... ....A..A. .....C.. AC.G..... ....C..... .....C.C ..G.....
HPCJK046E2 ..A..... ....C.. ....T.. ....A.....C. .G..CA.... C..G..... ..C..T..A. ....T..C.. C.AGAC....
HPCK3A ..A...G... ....G..AC.. ....C. T..A.....T ..A..... .G.....C.. AC.G..AA.. ....G..... .AG...A...
HCV4APOLY ....C... ....TC. ....G ....T.. A.....C..A ....G..... A.....A
HCV1480 ....A. ....TC. .....A...G ..A..... ....C..T.. AC.G..C... .....T.... ....G..C.. A...AC....
D84262 .....C ....AC.. ....C..A.. A.....A... ....A..C. .G..CA.... ...G..C..A ....A.... ....G..C.C AT.TC.....
D84263 ..C.....C ....CC.G ....T ..A.....C. .G..CA.... ...G..C... ..C..A..A. ....G..C.. C.AGAC....
D84264 ..C..... .A..... ....C... ..A.....C. ....CA.... ...G..... ..C..A..A. ....G..C.A .T.TC.....
D84265 ....C.T ....A.....C. .G...A.... C..G..C... ..C..A..A. ....G..C.A ...GATA...
HCV12083 ....A .....CC.G ....T ....A..C. .G..CA.... C..G..C..A .....A..A. ....G..C.A ....C.....
HPCJK049E1 .....G.AC .A..A..... .....C.AG ..... .T..AGT.C. C..... ....TA. ...G..C..A ..C..... G...G..C.. AA.....


220
339
AF011751 AGGACCTGGG CTCAGCCCGG GTACCCTTGG CCCCTCTATG GCAATGAGGG TTGCGGGTGG GCGGGATGGC TCCTGTCTCC CCGTGGCTCT CGGCCTAGCT GGGGCCCCAC AGACCCCCGG
AF064490 C..T..... G...A..... ....T.... C..... CCT..... ..A..G...T .G..C..C.. ...A..... .....AT. .....A T.....
AF511948 ....C..... C.....T... ....C..... ....
AF511950 ....T.. .... C..... ....
AY051292 ...T..... .G..... ....C..... C..T..... ..A..... ..T....C.. ...C.....C .....TC .....T. T.....
D50409 .A.T..... G.AG...A.. A..... .....G..CC .G..... CCT...C... ..A..... .....C.. ...C..T... ..C...TCA. .....G T.....
HEC278830 .... C..... ....C.. ....T.. .....C.
HPCCGS ...T..... .G..... ....C..... C..T..... ....C.. ...C..... .....T. .....TT. T.....
HPCHCJ1 .... C..... ....T. ..... G.....
HPCPLYPRE .... C..... ....
AB030907 .A.T..... GAA....A.. A..T..... ..T..G.... .A..C..... .....T... ..A..T.... .....C.. T..C..G... ..T....CT. .....G C.....
AB031663 .A.T..... GA.GT..A.. A.....C... .....G.... .G..... CCT..... ..A..... .....C.. ...A..... ..C...TCT. ....
AB047639 .A.G..... GAA.A..A.. TCG...C... .....A.... .G..... ACT...C... ..A..... .....C.. ...A..... ..C..CTC.. ..... T.....
AB047640 .A.T..... GGA....A.. A.....C... .....G.... .G..... GCT...T... ..A..... .....C.. T..A..T..C ..T..CTCA. .....A T.....
AB047641 .A.T..... GAA.A..A.. A..... .....G.... .G..... GCT...C... ..A..... .....C.. ...A..T... ..T..CTCT. .....A T.....
AB047642 .A.T..... GAAGA..A.. A..... .....A.... .G..... ACT...C... ..A..... .....C.. ...A..T..C ..C..CTCT. .....A T.....
AB047643 .A.T..... G.A.A..A.. A.....C... .....A.... .G..... GCT...C... ..A..... .....C.. ...G..T... ..T..CTCT. ....T...A T.....
AB047644 .A.T..... GAA.A..A.. A.....C... .....A.... .G..... GCT...C... ..A..... .....C.. ...A..T..C ..T..CTCT. .....T.. T.....
AB047645 .A.T..... GAA....A.. A..... .....A.... .G..C..... ACT...C... ..A..... .....C.. ...A..T..C ..T..CTCT. ..... T.....
AF169002 ...T..... GAA.A..A.. A.....C... .....A.... .G..... ACT...C... ..A..... .....C.. ...A..T..C ..T..CTCT. .....A T.....
AF169003 .A.T..... GAACA..A.. A..... .....A.... .G..... GCT...C... ..A..... .T.....C.. ...A..T..C ..T..C.CT. ..... T.....
AF169004 .A.T..... GAA.A..A.. A.....C... .....A.... .G..... ACT...C... ..A..G.... .....C.. ...A..Y..C ..T..CTCT. ..... T.....
AF169005 .A.T..... GAA.A..A.. A.....C... .....G.... .G..... ACT...C... ..A..G.... .....C.. ...A..T..C ..T..CTC.. ..... T.....
AF177036 .AAT..... GAA.A..A.. A.....C... .....A..C. .G..... ACT...C... ..A..... .....C.. ...A..T..C ..T..CTCT. .....A T.....
AF238481 .A.T..... GAA.A..A.. A.....C... .....A..C. .G..... ACT...C... ..A..... .....C.. ...A..T..C ..T..CTCT. .....A T.....
AF238482 .A.T..... GAA.A..A.. A.....C... .....A.... .G..... ACT...C... ..A..... .....C.. ...A..T..C ..T..CTCT. ..... T.....
AF238483 .A.T..... GAA.A..A.. A.....C... .....A.... .G..... ACT...C... ..A..... .....C.. ...A..T..C ..T..CTCT. ..... T.....
AF238484 .A.T..... GAA.A..A.. A.....C... .....A.... .G.....A.. ACT.T.C... ....C.. ...A..T..C ..T..CTCT. ..... T.....
AF238485 .A.T..... GAA.A..A.. A.....C... .....A..C. .G..C..... ACT...C... ..A..... .....C.. ...A..T... ..T..CTCT. .....A T.....
AF238486 .A.T..... GAA....A.. A..T..... .....G..C. .A..C..... .....C... ..A..T.... .....C.. ...C..G... ..T...CAT. ..... T.....
HPCJ8G .A.T..... GAA....A.. A..T..... .....G..C. .A..C..... .....C... .....T.... .....C.. ...C..G... ..T....CT. ..... C.....
HPCPOLP .AAT..... GAA.A..A.. A.....C... .....A..C. .G..... ACT...C... ..A..... .....C.. ...A..T..C ..T..CTCT. .....A T.....
AF046866 C..T..... .... .T..... C..... ..A..G.... .....C.. G..C.....C ..T..ATCT. .....A.A T.....
HCVCENS1 C..T..... .... .T..C..... C..... ..A..G.... .....C.. G..C..... ..T..ATCT. .....AA.A T.....
HPCEGS C..T..... .... .T..C..... C..... ..A..G.... .....C.. A..C.....C ..T..ATC.. .....A.A T.....
HPCFG C..T..... .C..... ....C. .G..... C..T..T... ..A..... ..T..C.C.. ...C..T... ..C..C.... .....AA.A T.....
HPCJK046E2 C.TG..... G.....T.. ....G.A... .....G.... .G..... C..... ..A..G.... .....C.. ...C..... ..C..A.CG. .....A.A C.....
HPCK3A C..T..... .C.....T.. .... .T..... C..... ..A..G.... .....C.. A..C.....C ..T..A.AT. ...C...A.A T.....
HCV4APOLY ...T..... .A..A..A.. A..T..A... ..T..T..C. .T..... ...T..... ..A..... ..T....C.. ....A..GTCT. .....A.A T..T.....C
HCV1480 C..T..... G...A..... ....T..C. C.....A.. CCT..... ..A..G...T .G..C..C.. ...AA..... .....AT. .....A T.....
D84262 C..... G.....T.. C.....C... .....G.... .G..... C..... ..A..C.... .TA..... ...C..... ..C...TCT. .....A.A C.....
D84263 C.T..... .A.....T.. ...T..C... .....C. .G..... C.....C... AT..... .....C.. .....C ..C..GCA.. .....T.A T.....
D84264 C..CA..... .....T.. T..T..C... .....T.... .T..C..... C..... ..A..G.... ..T..... ....T..C.A.. .....A T.....
D84265 C.T.G..... G...A..... C..... .....T.... ....C..... C.....A... ....C.. ...C..A... .....AT. .....A.A T.....
HCV12083 ...CA..... ..... A..... ..T..T.... .A.GC..A.. C..T..... ..A..T.... .....C.. ...C.....C .....ACA.. .....A T.....
HPCJK049E1 C..T..... ....A.... .G..... C..... ..A..G.... .....C.. G..C..... ..C..ATCG. .....A.A T.....


340
459
AF011751 CGTAGGTCGC GCAATTTGGG TAAGGTCATC GATACCCTTA CGTGCGGCTT CGCCGACCTC ATGGGGTACA TACCGCTCGT CGGCGCCCCT CTTGGAGGCG CTGCCAGGGC CCTGGCGCAT
AF064490 ..G.AA.... ....C..... ....
AF511948 .....T...
AF511950 .....C.... ...A.....
AY051292 ..C.....A.
D50409 .A..AA.... ....C..... C.....
HEC278830 ..C.....
HPCCGS ..G.....A.
HPCHCJ1
HPCPLYPRE
AB030907 .A...A..A. C.G..... ..C...A.C. ....T..T.. T.....T... .C..TG.... T..... G.C..... TC.....A.. T.....A..C
AB031663 .A...A..A. .....C.... C..... ....T..T.. T.....T... ....CG.... .....G G....C.... T......A.. A..C.....
AB047639 .A..... ....CG.... ...A..... ..C.....A. ....T..... T..... .C..CG.... A.....G ...A.T.... .C.....A.. TG.C.....C
AB047640 .A..... .....G.... .....A..... T..... T.....G .C..CG.... A.....G ...A.T.... T.....T.. T..C.....C
AB047641 .AC..... ....CG.... ...A..... ..C.....A. .....C.... TC.....A.. T..C.....
AB047642 .A..A.... ....CG.... .....A..... T..... .C..CG.... A.....G .....T.... T......A.. T..C.....C
AB047643 .A..... ....CG.... ....T...A. .....C. T..... .....T. .C..CG.... G.....G .....T.... TC.....A.. ...C.....
AB047644 .A..... .....G.... .....A.A. .A..... T..... .C..CG.... A.....G .....T.... T.....A.. T..C.....C
AB047645 .A..... ....CG.... .....A..... T..... .C..CG.... A.....A .....T.... T......A.. T..C.....C
AF169002 .A..... .....G.... ...A..... ..C.....A. .....Y.... TC....CA.. T..C.....C
AF169003 .A..... .....G.... .....A..... T..... .C..CG.... A.....G ..C..T.... TC.....A.. T..C.....C
AF169004 .A..... ....CG.... .....A..... T..... .T..CG.... A.....G .....T.... TC.....A.. T..C.....
AF169005 .A..... ....CG.... ..C.....A. .....C..CG.... A.....G .....T.... TC.....A.. T..C.....C
AF177036 .A..... ....CG.... .....A..... T..... .C..TG.... G.....G ..C..C.... TC.....A.. T..C.....
AF238481 .A..... ....CG.... .....A..... T..... .C..CG.... A.....G .....T.... TC.....A.. T..C.....C
AF238482 .A..... ....CG.... .....A..... ....T..... T..... .C..CG.... A.....A .....T.... T......A.. T..C.....C
AF238483 .A..... ....CG.... .....A..... T..... .C..CG.... A.....G .....T.... TC.....AA. T..C.....C
AF238484 .A..... .....G.... ...A..... T.....A. .....T..... .T..CG.... AA.....G .....T.... TC.....A.. T..C.....C
AF238485 .A..... .T..CG.... .....A..... .....C..CG.... A.....G .....T.... T......A.. T..C.....
AF238486 .A...A..A. C.GA..... .....A... ....T..T.. T..... .....T. .C..TG.... T.....G G.C..... TC.....A.. T.....A..C
HPCJ8G .A...A..A. C.GA..... .....A... ....T..T.. T..... .C..TG.... T.....G G..... TC.....A.. T.....A..C
HPCPOLP .A.....C. ....CG.... .....A..... T..... .C..TG.... A.....G ..C..C.... TC.....A.. T..C.....
AF046866 ..G.....C. ...A..... ....T..A.. .C..... .....C G.G..... TC..A..A.. ...C.....
HCVCENS1 ..G.....C. ....C..... C..... .....A.. .C..... .....T..C G.A..... TC..A..A.. ...C.....
HPCEGS ..G.....C. ...A..... .....A..... .....A.. .C..... .....T... G.A..... TC..A..A.. ...C.....
HPCFG .....A..C. ....C..... .....A..... .....A.. T..... .T.....A. ..... G.A..G.... TC..A..A.. ...C..T...
HPCJK046E2 .....T. ....C..... ..... .T.....C. ..... .C..CG..A. A....G...C T.A..G.... TC..AGCA.. T.....
HPCK3A ..AC....C. ...A..... .....C. .....A.. .....T... .C..... .....C ..A..... ....A..A.. ...C.....
HCV4APOLY G.G.....C. .....A. .C..... .....A.... .C..... A.....C G.G..TA... TC..... .....A...
HCV1480 ..G.AA..C. C..... ..G..... C.....A. .....A.. .C..C..... .....C..... A....G...C G....G.... TCT.A..... T..C..A..C
D84262 ...C....C. .T..... .....G. .A.....AC. ..... .T..CG.... A....GA..C T.G..T.... TC..GGCT.. G..C..A...
D84263 ..A.....C. ....C..... ...A..... ..... .C..... T..... ..... .C..TG.... ...T.....C ..A..G.... TC..TGCA.. G.....
D84264 ..G.....A. .T..C..... ....A. .C..T..... .... .T..CG.... G..G..G..G .....C..T. TC..GGCT.. .T.A..A..C
D84265 ..A.....T. ....C..... ....C. .....C. A.....T... ..... .C..TG..C. A..G.GG..G .....C.... T...GGCT.. .....
HCV12083 ...C.A..C. .G..... ....A. .C..T..G.. .....T... ...T..... .T..CG.... G.....G... T.G..C.... TC..GGCT.. G..T..A...
HPCJK049E1 .....A..C. ....C. .A.....A.. .... .T..... A..... G....G.... TC..A..A.. T..T..A...

```
           460                                                                 520
           ϕ                                                                   ϕ
AF011751   GGCGTCCGGG TTCTGGAAGA CGGCGTGAAC TATGCAACAG GGAACCTTCC TGGTTGCTCT T
AF064490   ..T..GA... .C..T..G.. ...G..A... .......... ....TT.A.. C.......... .
AF511948   ......A... .......... .......... .......... ....T..... .......... .
AF511950   .......... .......... .........T .......... .......... .......... .
AY051292   ..T..TA... ....A..... ......A..T ..C....... .......... .......... .
D50409     ..G..GA... .......G.. ...GA.A..T ........G. ....TT.G.. C........C .
HEC278830  .......... .......... .......... .......... .......... .......... .
HPCCGS     ..T..TA.A. .C........ .........T ..C....... .......C.. C.......... .
HPCHCJ1    .......... .......... .......... .......... .......... .......... .
HPCPLYPRE  .......... .......... .......... .......... .......... .......... .
AB030907   ..T..TA... .C........ ...GA.A..T ........GA ....TT.A.. C.......... .
AB031663   .....GA... .C.....G.. ...GA.A... .......... ....T..A.. C.......... .
AB047639   .....GA.A. .C.....G.. ...G..T..T .......... .......A.. C....T.C.C .
AB047640   .....GA.A. .C.....G.. ...G..T..T .T........ .....T.A.. .........C .
AB047641   .....GA.A. .C.....G.. ...G..T..T .T........ .....T.G.. .......... .
AB047642   .....GA.A. .CT....G.. ...G..T... .......... .......A.. .......... .
AB047643   .....GA.A. .C........ ...GTGT..T .......... .....T.A.. .........C .
AB047644   .....GA.A. .C.....G.. ...G..C..T .T........ .....T.A.. .........C .
AB047645   .....GA.A. .C.....G.. ...G..T..T .......... .......A.. .........C .
AF169002   .....GA.A. C......G.. ...GA.T..T .T........ .....Y.A.. .........C .
AF169003   .....GA.A. .C.....G.. ...G..T..T ..C....... .....T.A.. C........C .
AF169004   ..T..GA.A. .C.....G.. ...G..T..T .......... .....T.A.. .........C .
AF169005   .....GA.A. .C.....G.. ...G..T..T .......... .....T.A.. .........C .
AF177036   .....GA.A. .C.....G.. ...G..T..T .T........ .....T.A.. C........C .
AF238481   .....GA.A. .C.....G.. ...G..T..T .T........ .....T.A.. .........C .
AF238482   .....GA.A. .C.....G.. ...G..T..T .......... .....T.A.. .........C .
AF238483   .....GA.A. .C.....G.. ...G..T..T .......... .....T.A.. .........C .
AF238484   ..T..GA.A. .C.....G.. T..G..T..T .......... .....T.A.. .........C .
AF238485   .....GA... .......G.. ...G..C..T ..C....... ....T..A.. .........C .
AF238486   ..T..TA... .C.....G.. ...GA.A..T .......... ....TT.A.. C........C .
HPCJ8G     ..T..TA... .C.....G.. ...GA.A..T ..C....... ....TT.A.. C.......... .
HPCPOLP    .....GA.A. .C.....G.. ...G..T..T .T........ .....T.A.. C........C .
AF046866   .....GA... CC..T..... ...GA.A..T .TC..G.... .....T.G.. C........C .
HCVCENS1   .....GA... CC..T..G.. ...GA.A..T .TC....... .....T.G.. C........C .
HPCEGS     .....GA... CC..T..... ...GA.A..T .TC....... .....T.G.. C........C .
HPCFG      ..T..GA.A. CA..T..G.. ...A...... ..C....... ....T..... .........C .
HPCJK046E2 ......A... C.G.T..G.. ...G..A..T .......... ....T..... C........C .
HPCK3A     .....GA.A. CC..T..... ...GA.A..T .T........ .....T.G.. C........C .
HCV4APOLY  ..T...A... C.T....G.. ...GA.C..T .......... ....T..C.. C........C .
HCV1480    ..T..GAA.. ....T..... ...GA.A... .......... ....T..A.. C......C.. .
D84262     ..T..GA... CAA.C..G.. ...GA.T... .......... ....T..... .......... .
D84263     ......A... CCG.A..G.. ...GA.T... ..C....... .......... C........C .
D84264     ..T..GA... CAA.C..G.. ...GA.C... ..C....... ....T..... .........C .
D84265     .....TA... CAA.C..G.. ...G..C..T .......... ....T..... .......... .
HCV12083   ..T..GA... CAA.C..... ...GA.C..T .......... .A..T..... C.......... .
HPCJK049E1 ..T..T..A. C...T..G.. ...AA.C..T .TC....... ....TT.A.. .......... .
```

```
            8170                                                
```

```
          8290                                                
```

```
          8410                                                
```

```
          8530                                                
```

```
          8650                                                                                                                  8769
          ϑ                                                                                                                     ϑ
AF011751  CTGTGGGCGA GGATGATACT GATGACCCAT TTCTTTAGCG TCCTCATAGC CAGGGATCAG CTTGAACAGG CTCTTAACTG TGAGATCTAC GGAGCCTGCT ACTCCATAGA ACCACTGGAT
AF064490  ........TC ....AG.G.. ......G..C ..T..C.... .A..GCAGAG .CA...G..A ..G...A.A. .C..AGCA.T ...A..G..T ..GT.AGT.. ....TG.CAC T..G......
AF511948  .......... .......... .........C ..A..... .......... .........A ....T..... .C...G.T.. C..A..... .......... .......... ..........
AF511950  .......... .......... ......T..C .......... ....T..... .......... .......... .C...G.TG. C..A..... .......... .......... ..........
AY051292  .......TAC .....G.C.. C......... ..T..CTC.A .A........ TCA...G..C ....G.A... ....AG.T.. ...A.....T ......GTAC .....G.CC. ...GT....C
D50409    G.T....T.C .C...G.C.. A.....A..C .....CTC.. ..T.G..G.. .CA....ACC ..G..C.... A...C....T C..A..G... .....AGTG. ....AG.CAG T..C.....C
HEC278830 T......... .......... .........C .....C..T. .......G.. .......... ..C..... .C...G.T.. ...A..... .......... .......... .........C
HPCCGS    .......TAC .....G.C.. T......... ..T..CTC.A .A........ .CA...A..C ......A... ....AG.T.. ...A.....T ..G...GTGC .....G.CC. ...GT.A...
HPCHCJ1   T......... .......... .........C .....C..T. .......... .......... .......... .C...G.T.. C..A..... .......... .......... ..........
HPCPLYPRE .......... .......... .......... .......... ....T..... ......C... .......... .C..CG.T.. C......... ..G..... .......... ......T...
AB030907  A.C....TCC .....G.CA. A........C ..T..CTC.A .A..AC.G.. .CA...CACT ..GA.C...A A...C..T.T C.....G... ..G..AGTA. ....GG.CA. T.....A..C
AB031663  A.A....T.C .C...G.C.. C.....A..C ..T..CTC.. ....A..G.. .CA...AACC ..G..T..A. AC..A....T ...A..G... .....AGTG. .T..TG.GA. C..T.....C
AB047639  A.A....TTC .C...G.C.. A.....A..C .....CTC.A .T.....G.T .CAA..CACC ..G..C...A AC..C....T ......G..T ...T.AGTA. .....G.GA. T..TT....C
AB047640  A.A....TTC .C...G.C.. ......A..C .....CTC.A .T.....G.. .CA...CACC ..A..C...A AC.......T ...A..G... ...T.GGTG. .....G.GAG T..T.....C
AB047641  A.A.....TC .C...G.C.. ......A..C .....CTC.A .......G.. .CAA..CACT ..G..C...A AC..C..T.T ......G..T .....GGTG. .....G..AG T..TT....C
AB047642  A.A....TTC .C...G.C.. ......A... ......TC.A .T.....G.. .CAA..CACC ..G..C...A AC..C....T ...A..G... ...T.GGTG. .....G.GAG T..TT....C
AB047643  A.A....TTC .C...G.C.. ......A..C ....A.TC.A ....T..G.. .CAA..CACC ..G..C...A GC.......T ......G... .....GGTG. ....AG.GAG T..TT....C
AB047644  A.A....TTC .....G.C.. ......A..C .....CTC.A .T..T..G.. .CAA..CACC ..G..C...A AC..C....T ...A..G... ...T.GGTG. .....G.GAG T..TT....C
AB047645  A.A....TTC .C...G.C.. ......A..C .....CTC.A .T..T..G.. TCAA..CACC ..G..C...A AC..C..T.T ......G... ...T.GGTG. .....G.GAG T..TT....C
AF169002  A.A....TTC .C...G.C.. ......A..C ..T.ACTC.A .T.....G.. .CAA..CACC ..G..C...A AC.......T ......G... .....GGTG. .....G.GAG T..CT....C
AF169003  A.A....TTC .T...G.C.. ......A..C .....CTC.A .T.....G.. .CAA..CACC ..G..C...A AC..C....T ...A..G... ...T.GGTG. .....G.GAG T..TTC...C
AF169004  A.A....TTC .T...G.C.. ......A..C .....CTC.A .T.....G.. .CAA..CACC ..G..C...A AC..C....T ...A..G... ...T.GGTG. .....G.GAG T..TTC...C
AF169005  A.A....TKC .C...G.C.. ......A..C .....CTC.A .T.....G.. TCAA..CACC ..A..C...A AC..C....T ......G... ..GK.GGT.. .....G.GAG T..TT....C
AF177036  A.A.....TC .C...G.C.. ......A..C .....CTC.A .T.....G.. TCAA..CAC. ..G..C...A AC..C....T ......G... .....GGTG. .....G.GAG T..CT....C
AF238481  A.A....TTC .A...G.C.. ......A... .....CTC.A .T.....G.. .CAA..CACT ..G..C...A AC.......T ...A..G... ...T.GGTG. .T...G.GAG C..TT....C
AF238482  G.A....TTC .C...G.C.. ......A..C .....CTC.A .T.....G.. .CAA..CACC ..G..C...A AC..C....T ...A..G... ...T.AGTG. ....TG.GAG T..TT....C
AF238483  A.A.....TC .C...G.C.. ......A..C .....CTC.A .T.....G.. TCAA..CACC ..G..C...A A...C....T ......G... ...T.GGTG. .....G.GAG T..TT....C
AF238484  G.......TC .C...G.C.. ......A..C ..T..CTC.A .T.....G.. .CAA..CACC ..G..C...A A...C....T ......G... ...T.GGTG. .....G.GAG T..TT....C
AF238485  A.A....TCC .C...G.C.. ......A... ..T..CTC.A ....T..G.. .CAA..CACT ..G..C.... AC.....T.T ...A..G... ...T.GGTG. .....G.GAG T..T.....C
AF238486  A.C....TCC .....G.CA. A........C ......TC.A .A..AT.G.. .CA...CACT ..GA.C..AA A...C..T.T ......G..T ..G..AGTA. .T..GG.CA. T...T.A..C
HPCJ8G    A.C....TCC .....G.CA. A.....T..C .....CTC.A .A..AT.G.. .CA...CACT ..GA.C..AA A...C..T.T ......G... ..G..AGTA. ....GG.CA. T...T.A..C
HPCPOLP   A.A.....TC .C...G.C.. ......A..C .....CTC.A .T.....G.. .CAA...ACT ..G..C...A AC..C....T ......G... .....GGTG. .....G.GAG T..CT....C
AF046866  A.C....T.C .C...G..A. ......A..C ..T..CTC.A .A...CA.T. .CA...GATA .....T.GAC .C...G...T C..A..G... ..G...ACT. ....GG.CAC G..G......
HCVCENS1  A.C....T.C .C...G.GA. ......A..C ..T..CTC.A .A...CA.T. .CA...GATA .....T.GAC .C...G...T ...A..G... ..G...ACT. ....TG.CAC T..G......
HPCEGS    A.C....T.C .C...G.GA. ......A..C ..T..CTC.A .A...CA.T. .CA...GATA .....T.GCC .C...G...T ...A..G... ..G...ACT. .T..TG.CAC T..G......
HPCFG     A.T....TCC .C...G.GT. ...A.....C ......TC.A .A..GCA... .CA...G... ......AGA. .C..AG...T ...A..G... ..G..TAC.. ....AG.CAC T.........C
HPCJK046E2 A.A....T.C .T...G.G.. ......G..C ......G.GA .A...CAGC. TCAA..G... .CC.CA.A. .A..CG...T ...C..G... ..G.T.ACG. ..AAT..CAC T.....T...
HPCK3A    A.C....T.C .C...G..A. ......G... ..T..CTC.A .A...CA.T. .CA...GATA .....T.GAC .C...G...T ...A..G... ..G...ACT. ....TG.CAC T..G......
HCV4APOLY A.A....T.C .C......T. .........C ..T..CTCAA .A...CAGAG .CA...AGCC .....GA.A. .A..CG...T C..T..G... ....T.AC.. ....T..CAC T..G......
HCV1480   T.......TC ....AG.G.. ......G..C ..T..C.... .T..ACAGAG .CA...A... ..G...A.AA .C..AGCA.T ...A..G..T ..GT.AGTT. .T..TG.TAC T..G......
D84262    A.A....TAC .C...G.C.. C......... .....CTC.A .....CA.TG .CA...G... T.G..GGC.. .A.....T.T ...C..G..T ..G.TAAC.. .....G.CAC T.....C...
D84263    A.T....TAC .T...G.GT. A.....T... ......GCAA .A...CA.T. .CA...AATA ..CC.TA.A. .A...G...T C..C..G... ....T.ACA. .T...G.CAC T.....T..C
D84264    A.C....TAC .T...G.G.. C........C .....CCAAA .A...CA... GCA...G..A ..A..TA... TG..AG...T C..T..G... ....T.ACA. .T..AG.TAG T.....CC.A
D84265    A.T....TAC .T...G.... C......... ......CAAA .A...CA... ACA...GACT T.A..TAG.. .A...G...T C..C..... ....T.ACG. .T..A..CAC T..G..T...
HCV12083  A.A....T.C .T...G.T.. .....A.... ..T..CTC.A .....CAGT. GCAA..A..A T.G..GA.A. .AT.CG...T ...C..... ....TGAC.. .T...G.TTC T.....T...
HPCJK049E1 A.C....T.C .C...G.GA. T......... ..T..CTC.A .A..GCA... TCA...G..A ..A...A.A. .C..CG...T C.....G... ..G..TGT.. .T..TG.CAC T.........C
```

```
          8770                                                
```

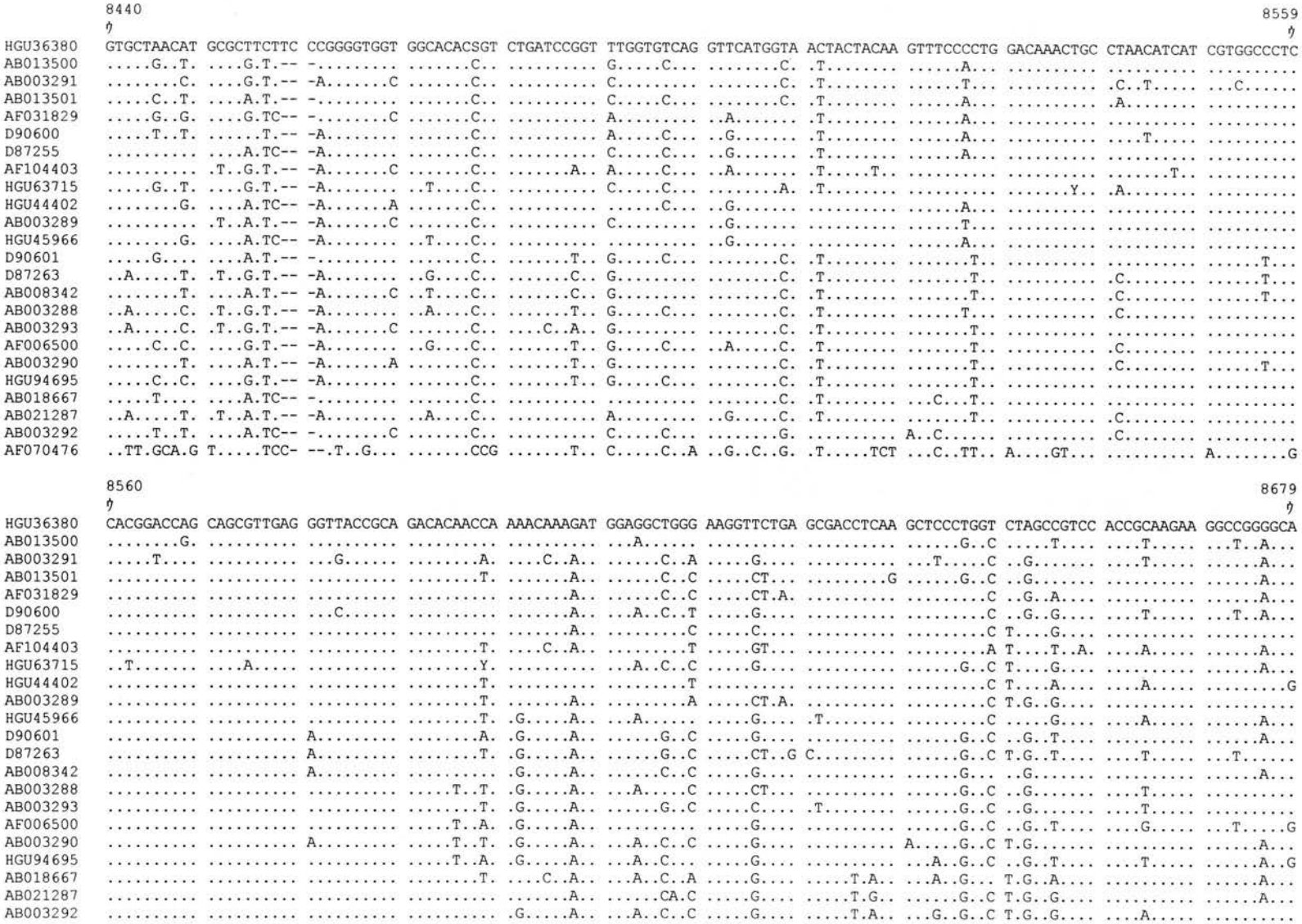
```
          8890                                                
```

```
                9010                                                
```

```
            9130                                                                                                9200
            ϕ                                                                                                   ϕ
AF011751    CTCGCTGCAG GGGTAGGCAT CTACCTCCTC CCCAACCGAT GAAGGTTGGG GTAAACACTC CGGCCTCTTA A
AF064490    ..TA.C.TG. .......... ..TTT.A... ..TGCT..R. A.GCTGG.A. C......... .ATAACTCCT G
AF511948    .......... .......... .......... .......... .......... .......... .......... G
AF511950    ..T....... .......... .......... .......... .......... .......... .......... G
AY051292    ..TT.A.T.. .......... ...T.....T .......... AG.C.G?T.. .C..C..... ...GTCT... G
D50409      ..TTGC.T.. .......... ..T...A... ...GCT..G. AG..CGGCAC A.TT.GCTA. ACT..ATAGC T
HEC278830   ..T....... .....T.... .......... .......... ...C.GG.A. C......... .A.G.CAA.. G
HPCCGS      ..TT.A.TG. .......... T........T ........G. AG.C.G.T.. .C..C..... .A.G.CT... G
HPCHCJ1     ..T....... .......... .......... .......... A......... .......... .......... G
HPCPLYPRE   ..T....... .......... .......... .......... .......... .......... ......???? ?
AB030907    ..TAGC.T.. .A........ ..TTT.A... ...GCT..G. AG..CGGCAA ACCCTAG..A .ACT.CA.AG C
AB031663    ...AGC.T.. .......... ..TTT.A... ...GCT..G. AG..CGGCAC ACCTTAG..A .ACT.CA.AG C
AB047639    ..TTTC.T.. ........C. ..T...A... ...GCT..G. AG..CGGCAC AC.CTAGG.A .ACT.CA.AG C
AB047640    ..TT...T.. ........C. ..T...A... ...GCT.... AG..CGGCAC AC.TTAG..A .ACT.CA.AG C
AB047641    ..TT.C.T.. ........C. ..T...A... ...GCT..G. AG..CGGCAC AC.TTAG..A .ACT.CA.AG C
AB047642    ..TT...T.. ........C. ..TTT.A... ...GCT..G. AG..CGGCAT AC.TTAG..A .ACT.CA.AG C
AB047643    ..TTG..T.. .......... ..TT..A... ...GCT..G. AG..CGGCAC AC.TTAG..A .ACT.CA.AG C
AB047644    ..TTTC.T.. ........C. .CTT..A... ...GCT..G. AG..CGGCAC AC.TTAG..A .ACT.CA.AG C
AB047645    ..TTTC.T.. ........C. ..TTT.A... ...GCT..G. AG..CGGCAC AC.TTAG..A .ACT.CA.AG C
AF169002    ..TTTC.T.. ........C. ..T...A... ...GCT..G. AG..CGGCAC AC..TAG..A .ACT.CA.AG C
AF169003    ..TT.C.TG. .AC.....C. T.T...A... ...GCT..G. AG..CGGCAC AC.CTAG..A .ACT.CA.AG C
AF169004    ..TT.C.TG. .AC.....C. T.T...A... ...GCT.... AG..CGGCAC AC.TTAG..A .ACT.CA.AG C
AF169005    ..TCT..T.. ........C. T.T...A... ...GCT..G. AG..CGGCAC AC.TTAG..A .ACT.CAYAG C
AF177036    ..TTT..T.. ........C. T.T...A... ...GCT..G. AG..CGGCAC AC.TTAG..A .ACT.CA.AG C
AF238481    ..TT.C.T.. ........C. ..T...A... ...GCT.... AG..CGGCAC ?????????? ?????????? ?
AF238482    ..TTTC.T.. ........C. ..T...A... ...GCT..G. AG..CGGCAC ?????????? ?????????? ?
AF238483    ..TT...T.. ........C. ..TT..A... ...GCT..G. AG..CGGCAC ?????????? ?????????? ?
AF238484    ..TT.C.T.. ........C. ..T...A... ...GCT.... AG..CGGCAC ?????????? ?????????? ?
AF238485    ..TT.C.T.. ........C. ..T...A... ...GCT..G. AG..CGGCAC ?????????? ?????????? ?
AF238486    ...AGC.T.. .A........ ..TTT.A... ...GCT..G. AG..CGGCAA AC???????? ?????????? ?
HPCJ8G      ..TAGC.T.. .A........ ..TTT.A... ...GCT..G. AG..CGGCAA ACCCTAG..A .ACT.CA.AG C
HPCPOLP     ..TTT..T.. ........C. T.T...A... ...GCT..G. AG..CGGCAC AC.TTAG..A .ACT.CA.AG C
AF046866    ..AA.G.T.. .......... ..TT.....G ..AGCT..G. ..GCTGGTAA .AT....... .A???????? ?
HCVCENS1    ..AA.G.TG. .......... ..TT.....G ..AGCT..G. ..GCTGGTA. .AT....... .ATTTCT.?? ?
HPCEGS      ..AA.A.T.. .......... ..TT...T.G ..AGCT.... ..GCTGGTAA .AT....... .ATTTCT..T T
HPCFG       ..TA.C.T.. .......... T.T......A ..TGCT..G. ..GCTGGTA. ..T.....C. .AA..CTG.G T
HPCJK046E2  ..TA.A.T.. .......... ..TTT.T... ...CC..... A..C.GGC.. CCTC.AG.AA .ACT.CA..T G
HPCK3A      ..AA.A.T.. .......... ..T......G ..AGCA..G. ..GCTGGTA. .AT....... .ATT.CT..T T
HCV4APOLY   ..TA.A.T.. .......... ..T...G..G ..TGCT..G. AGGCAGCTTA AC.CT.CGA. .TTAGGG.CC C
HCV1480     ..TAGC.T.. .......... ..TTT.A... ...GCT..G. ..GCTGG.A. C......... .????????? ?
D84262      ...A...T.. .......... ..TTT.A... ...GCT..G. AG.C.AGC.. CCCT.A..AA .ACT.CA.A. C
D84263      ...A.C.T.. .......... ..TTT.G... ..TGCT..G. A.GC.GG.A. C.T..AG.AA .ACT.CA... G
D84264      ...T.C.T.. .......... ..TTT.A..A ...GCT..G. AG.C.GGTA. C.T....... .ATTTCTG.T T
D84265      ..GT.C.T.. .......... ..T...G... ..AGC...G. AGGCTGA.A. C........A .ACT.CG..T T
HCV12083    ..TA.C.T.. .......... ..TTT.G... ...GCT..G. AG???????? ?????????? ?????????? ?
HPCJK049E1  ..TA.G.T.. .......... ..T...GT.A ..AGC..... AGGCTGGTAA C...CACTC. ATTTT.AACT T
```

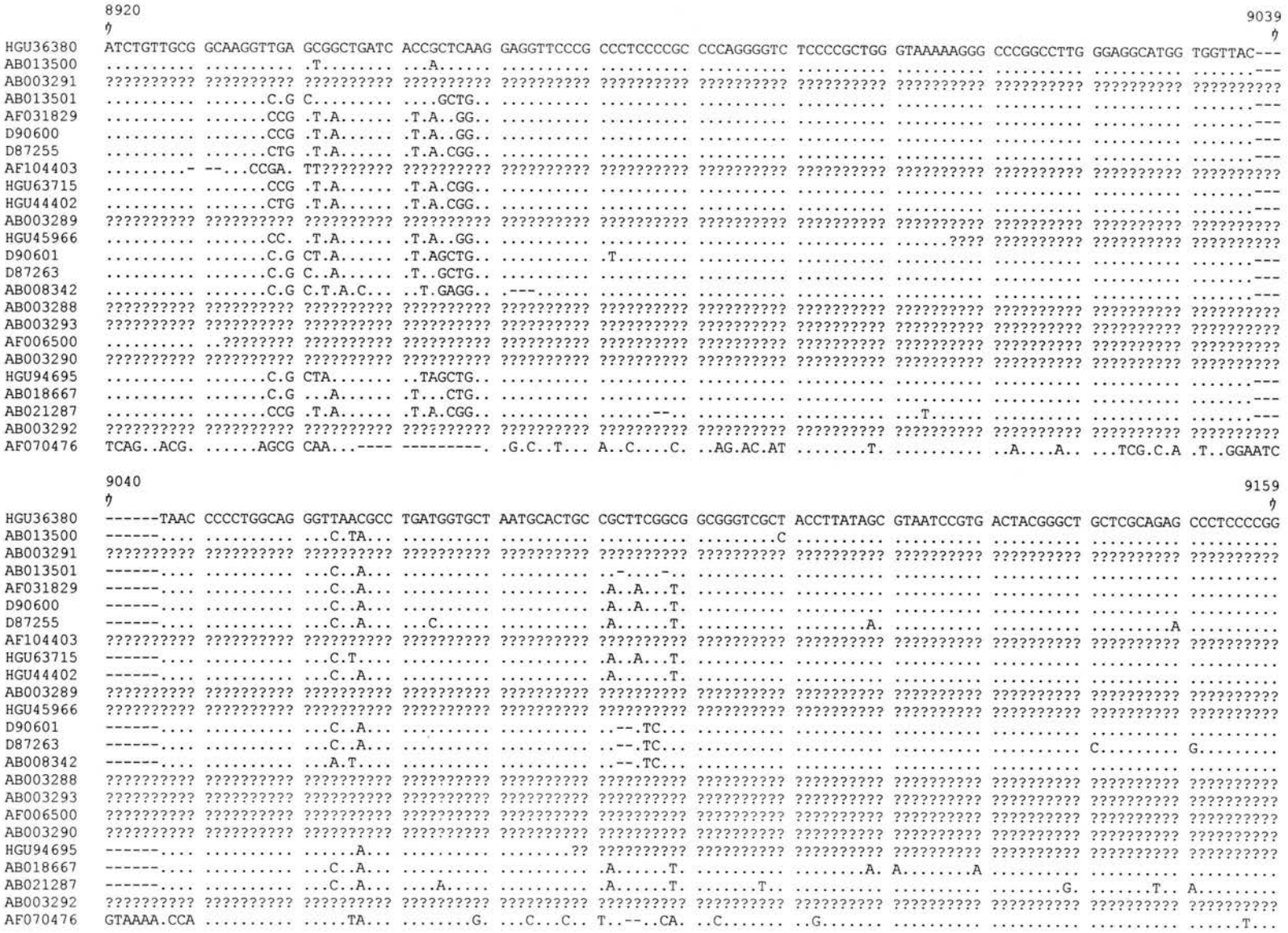
```
          8200                                                                                                                  8319
          ϑ                                                                                                                     ϑ
HGU36380  GCTTCGTACG GGTACGCGTG TGAGCCCTCG TATCACGCTT CACTGGACAC AGCCCCCTTC TGCTCCACTT GGCTCGCTGA GTGCAATGCG GATGGGRAAA GGCATTTCTT CCTGACCACG
AB013500  .....A.... ....T..... ...A.....A .....T..A. .......... G......... .......... ....T..C.. .........A ......A.GC .C........ .T........
AB003291  ..C....... .......... ...A...... .....T..A. ..T....T.. G......... .......... .......C.. .........T ..C...A..C .CTG...... ...T......
AB013501  .......... .A..T..... .......... .....T..A. .......... G......... .......... ....A..... .........A ..C...A.GC .C........ .........T
AF031829  ..GAGT.... ....T..... C.....G... .....T..A. ..T....... G......... .......... ....T..C.. .........A ......A.GC .C........ T.........
D90600    ...AGC..T. .......... C.....T... .....T..A. ..T....... G......... .......... ....T..C.. .........A ......A.GC .C........ ..........
D87255    ..AAGC..T. .......... C.....T..A .....T..A. ..T....... G......... .......... ....T..... ...T.....A ......A.GC .C........ ..........
AF104403  ...AGC..T. .......... C.....T..A .....T..A. ..T....... G......... .......... ....T..C.. ...T.....A ......A.GC .C........ ..........
HGU63715  ...AGC..T. ....T..A.. C.....T... .....T..A. .......... G......... .......... .......C.. ...T.....A ......A.GC .C........ ......A...
HGU44402  ..GAGC..T. .......... C.........A .....T..A. ..T....... G......... .......... ....T..... .........A ......A.GC .C........ ..........
AB003289  ..GAGC.... .A........ C.........A .....T..A. ..T....... G......... .......... ....T..... .........A ..C...A.GC .C........ ......T...
HGU45966  ..GAGC..T. ....T..T.. C.....T... .....T..A. .......... G......... .......... .......... ......C..A ......A..C .C........ ..........
D90601    .......... .......... ......T... .....T..A. ..T....... ...T...... .......... .....A..C. .........T ......A.GC .T........ T..T......
D87263    .....A.... .......... .........A .....T..A. .......... G......... ........C. ....T..C.. .........A ......A.GC .C........ ...T......
AB008342  .....A..T. .......... .........A .....T..A. ..T....... ...T...... .......... ....T..... .........T ......A.GC .T........ T..T..T...
AB003288  .GC..C..T. .......... .........A .....T..A. .TT.A..... G......... .......... ....A..C.. .........A ......A.GC .C........ ..........
AB003293  .....A..T. .......... .........A .....T..A. .......... C..T....... .......... ....A..C.. .........A ..C...A.GC .C........ ..........
AF006500  .......... .A........ .........A .....T..A. ..T.....G. G.....T... .......... ....T..... .........A ......A.GC .C........ ..........
AB003290  ..C..C.... ....T..... ...A.....A .....T..A. ..T....... G......... .......... ....T..C.. ...T.....T ..C...A..C .C........ ...T......
HGU94695  .GG..T.... .......A.. .......... .....T..A. ..T....... G......... .......... ....A..... ...T.....A ..C...AGGC .C........ ......A...
AB018667  .......... ....T..... .........A .....T..A. .......... G......... .......... ....T..... .........A ......A.GC .C........ ...A......
AB021287  ...AGC..T. .......... C......... .....T..A. ..T....... G......... .......... ....T..... ...T.....A ......A.GC .C........ ..........
AB003292  .......... ....T..... .........A .....T..A. ..T....... G......... .......... ....T..... ...T.....A ..C...AG.C .C........ T.........
AF070476  CGC....... .....AAT.. ...A..A..C .....T..C. .C..T..... T..T..G..T .....G..C. ....G..A.. ...T.....C A....TAGT. ....C..T.. .T....T..T

          8320                                                                                                                  8439
          ϑ                                                                                                                     ϑ
HGU36380  GACTTTCGGA GACCACTCGC TCGCATGTCG AGCGAGTACA GTGACCCTAT GGCTTCGGCC ATTGGTTACA TTCTCCTCTA CCCCTGGCRT CCCATCACAC GGTGGGTCAT CATCCCGCAT
AB013500  .......... .G..G..... .......... .......... .......A.. .........T ..A....... .C.....T.. ........AC .......... .......T.. ......T..C
AB003291  ..Y....... .G........ .......... .......... .......A.. ...G..A... ..C....... .C.....T.. ........A. .......... .......... ......C..C
AB013501  .......... .G........ .......... ..T.....T. .......A.. .......... .......... .C.....T.. ........A. .......... .......... ......T...
AF031829  .......... ....G..T.. .......... ........T. .......A.. .......... ..A....... .......T.. T.......A. .......... .......... ......T..C
D90600    .......... .G..G..... .......... .......... .......... .......... .......... .C.....A.. T.......AC .......... .......... ......T...
D87255    .......... .G..G..... .......... .......... .......G.. .......... .......... .C.....T.. T..A....AC .......... .......... ......T...
AF104403  .......... .G..GT.... .......... .......... .......G.. .........T ..G....... .C.....A.. ........A. .......... .......... ......T..C
HGU63715  .......... .G..G..... .......... ..T....... .......G.. .......... ..C....... .C.....T.. ........A. .......... .......... ...T..A..C
HGU44402  .....C.... .G..G..... .......... ..T.....T. .......G.. .........G ..C....... .C.....T.. T..T....AC .......... .......... ......T...
AB003289  .......... .G..G..... .......... ..T....... .......G.. .......... .......... .C.....A.. T.......A. .......... .......... ...T..T...
HGU45966  .......... .G..G..T.. .......... ........T. .......A.. .......... ..A....... .C.....G.. T.......A. .......... .......... ......T...
D90601    .......... .G...T.G.. .........A .......... .......A.. .......... .......... .C.....T.. ........AC .......... .......... ......C...
D87263    .......... .G..G..G.. .......... .......... .C..T..C.. .......... ..C..G.... .C.....T.. ........AC .......... .......... ......T..C
AB008342  .......... .G..G..G.. A.........A ..T....... .C.....G.. ......A..T .....G.... .C.....A.. ........AC .......... .......... ......A...
AB003288  .......... .G..G..G.. A.........A ..T....... .C.....A.. .......... ..C..G.... .C.....A.. ........AC .......... .......... ......T..C
AB003293  .......... .G..G..G.. A.........C .......... .......A.. .......... .......... .C.....T.. ........AC .......... .......... ......T..C
AF006500  .......... .G.....G.. .........A ..T....... .C.....C.. .......... .....G.... .C.....A.. ........A. .......... .......... ......T..C
AB003290  .......... .G..G..G.. A.........A ..T....... .C........ .......... .....G.... .C.....A.. ........AC .......... .......... ..........
HGU94695  .......... .G.....G.. C.........A .......... .C.....C.. .......... ..C..G.... .C.....A.. ........A. .......... .......... ......C..C
AB018667  .......... .......G.. ...T.....C ..T....... .......A.. .......... ..C....... ......CT.. T.......A. .......... .......... ..........
AB021287  .......... .G..G..... .......... .......... ....T..... .........A ..A....... .C.....A.. ........A. .......... .......... ......T...
AB003292  .......... .G.....A.. A..G.....T ..T....... .......... .......... .......... .C.....T.. ........AC .......... .......... ......C...
AF070476  ......A... ....TT.G.. AA.A.....C ..T.....TT CA.....A.. .........A .....G..TT G.T.G..... ...T....AC ..G..AGTCA .......TGC ...A..C...
```


8440
8559
HGU36380 GTGCTAACAT GCGCTTCTTC CCGGGGTGGT GGCACACSGT CTGATCCGGT TTGGTGTCAG GTTCATGGTA ACTACTACAA GTTTCCCCTG GACAAACTGC CTAACATCAT CGTGGCCCTC
AB013500 .....G..T. ....G.T.-- -..... ....C.. ..... G.....C... ....C. .T..... ......A... ....
AB003291 ....C. ....G.T.-- -A.....C ....C.. ..... C..... ....C. .T..... ......T... ..... .C..T..... ...C.....
AB013501 .....C..T. ....A.T.-- -..... ....C.. ..... C.....C... ....C. .T..... ......A... ..... .A..... .....
AF031829 .....G..G. ....G.TC-- -.....C ....C.. ..... A..... ..A..... .T..... ......A... ....
D90600 .....T..T. ......T.-- -A..... ....C.. ..... A.....C... ..G..... .T..... ......A... ....T..... .....
D87255 ....A.TC-- -A..... ....C.. ..... C.....C... ..G..... .T..... ......A... ....
AF104403 ..... .T..G.T.-- -A.....C ....C.. .....A.. A.....C... ..A..... .T.....T.. ....T.. .....
HGU63715 .....G..T. ....G.T.-- -A..... ..T....C.. ..... C.....C... ....A. .T..... ....Y.. .A..... .....
HGU44402 ....G. ....A.TC-- -A.....A ....C.. ....C... ..G..... ....A... ....
AB003289 ..... .T..A.T.-- -A.....C ....C.. ..... C..... ..G..... ....T... ....
HGU45966 ....G. ....A.TC-- -A..... ..T....C.. ....G..... ....A... ....
D90601 .....G.... ....A.T.-- -..... ....C.. .....T.. G.....C... ....C. .T..... ......T.. ....T...
D87263 ..A.....T. .T..G.T.-- -A..... ..G....C.. .....C.. G..... ....C. .T..... ......T.. ..... .C..... .....T...
AB008342 ....T. ....A.T.-- -A.....C ..T....C.. .....C.. G..... ....C. .T..... ......T.. ..... .C..... .....T...
AB003288 ..A.....C. .T..G.T.-- -A..... ..A....C.. .....T.. G.....C... ....C. .T..... ......T... ..... .C..... .....
AB003293 ..A.....C. .T..G.T.-- -A.....C ....C.. ....C..A.. G..... ....C. .T..... ......T.. ....
AF006500 .....C..C. ....G.T.-- -A..... ..G....C.. .....T.. G.....C... ..A.....C. .T..... ......T.. ....
AB003290 ....T. ....A.T.-- -A.....A ....C.. .....T.. G..... ....C. .T..... ......T.. ..... .C..... .....T...
HGU94695 .....C..C. ....G.T.-- -A..... ....C.. .....T.. G.....C... ....C. .T..... ......T.. ....
AB018667 .....T.... ....A.TC-- -..... ....C.. ....C. .T..... ...C...T.. ....
AB021287 ..A.....T. .T..A.T.-- -A..... ....A....C.. ..... A..... ..G.....C. .T..... ......T.. ..... .C..... .....
AB003292 .....T..T. ....A.TC-- -.....C ....C.. ..... C.....C... ....G. ..... A..C..... ..... .C..... .....
AF070476 ..TT.GCA.G T.....TCC- --.T..G... ....CCG .... C.....C..A ..G..C..G. .T.....TCT ...C..TT.. A....GT... ..... A.....G
8560
8679
HGU36380 CACGGACCAG CAGCGTTGAG GGTTACCGCA GACACAACCA AAACAAAGAT GGAGGCTGGG AAGGTTCTGA GCGACCTCAA GCTCCCTGGT CTAGCCGTCC ACCGCAAGAA GGCCGGGGCA
AB013500 ....G. ....A..... ....G..C .....T.... ....T..... ...T..A...
AB003291 .....T.... ....G..... ....A. ....C..A.. ......C..A .....G.... ....T.....C ..G..... ....T..... ......A...
AB013501 ....T. ....A.. ......C..C .....CT... ....G ......G..C ..G..... ....A...
AF031829 ....A.. ......C..C .....CT.A. ....C ..G..A.... ....A...
D90600 ....C..... ....A.. ...A..C..T .....G.... ....C ..G..G.... ....T..... ...T..A...
D87255 ....A.. .....C ....C.... ....C T....G.... ....
AF104403 ....T. ....C..A.. ....T ....GT... ....A T....T..A. ....A..... ......A...
HGU63715 ..T..... ....A..... ....Y. ....A..C..C .....G.... ....G..C T....G.... ....A...
HGU44402 ....T. ....T ....C T....A.... ....A..... ....G
AB003289 ....T. ....A.. ....A .....CT.A. ....C T.G..G.... ....
HGU45966 ....T. .G.....A.. ...A..... .....G.... .T..... ....C .....G.... ....A..... ......A...
D90601 ....A..... ....A. .G.....A.. ......G..C .....G.... ....G..C ..G..T.... ....A...
D87263 ....A..... ....T. .G.....A.. ......G..C .....CT..G C..... ....G..C T.G..T.... ....T..... ...T.....
AB008342 ....A..... .... .G.....A.. ......C..C .....G.... ....G... ..G..... ....A...
AB003288 ....T..T. .G.....A.. ...A.....C .....CT... ....G..C ..G..... ....T..... .....
AB003293 ....T. .G.....A.. ......G..C .....C.... .T..... ....G..C ..G..... ....T..... .....
AF006500 ....T..A. .G.....A.. ....G.... ....G..C ..G..T.... ....G..... ...T.....G
AB003290 ....A..... ....T..T. .G.....A.. ...A..C..C .....G.... .... A.....G..C T.G..... ....A...
HGU94695 ....T..A. .G.....A.. ...A..C... .....G.... ....A..G..C ..G..T.... ....T..... ......A..G
AB018667 ....T. ....C..A.. ...A..C..A .....G.... ....T.A.. ...A..G... T.G..A.... ....A...
AB021287 ....A.. ......CA.C .....G.... ....T.G.. ....G..C T.G..G.... ....A...
AB003292 .... .G.....A.. ...A..C..C .....G.... ....T.A.. ....G..G..C T.G..G.... ....A..... .....

```
AF070476  .....T.... .......A.. ...C..T... .......... ....C..A.. .......... .....G..AT C....A.G.. A.....G..C ..CT.T...T ....G..A.. ...T..T..G
          8680                                                                                                                 8799
          ϑ                                                                                                                    ϑ
HGU36380  TTGCGAACAC GCATGCTCCG GTCGCGCGGT TGGGCGGAGT TGGCTAGGGG CCTGTTGTGG CATCCAGGAC TCCGGCTTCC TCCCCCTGAG ATTGCTGGTA TCCCAGGGGG TTTCCCTCTG
AB013500  .....G.... .......... .........A .....A.... .....C.... .......... ........C. .......C.. .......... .........G .......... ..........
AB003291  C......... .......... .......... .........C .......... .......... ........C. .A.....C.. .......... ........A. .......A.. .......T..
AB013501  C....T..G. .......... .......... .........C .....C.... .......... .....T.... .......... C.....C... .......... ....G..... ...T..C...
AF031829  C....G.... .......... T......... .....T...C .......... .......... ........C. .A.....C.. .......... .......... ....G..... ...T..C..T
D90600    .......... .......... .......... .....T...C .......... .......... ........C. .......C.. C..T..C... .......... ....G..... ......C..T
D87255    .....G..T. .......... T......... .....C.... ....C..... .......... ........C. .G........ C.....C... .......... ....G..... ......C..T
AF104403  .....G.... .......... C......... .....A.... .......... .......... ........C. .A........ .......... .......... .......... G..T..C..C
HGU63715  C.......T. .......T.. .......... .....T.... .....C.... .......... ........C. .A.....C.. A..T..G... .......... ....C..... ......C...
HGU44402  .......... .......... C......... .....T.... .......... .T........ ........C. .A........ .......... .......... ....G..... .........C
AB003289  C.......T. .......... T......... .....C...C .......... .T........ ......T.C. .A........ C.....C... .......... ....G..A.. .........T
HGU45966  C......... .......T.. .......... .....C.... ....G..... .......... ........C. .......C.. .......... .......... ....G..... ......C..C
D90601    C.A..G.... .T........ .......... .....A...C .....C.... .......... ........C. .......C.. C.....C... .......... ....G..... ......C...
D87263    .....G..T. .......... .........A .....A...C .....C.... .T........ ........C. .......C.. ......C... .........G ....G..... ......C...
AB008342  .....C..G. .T.....T.. .......... .........C .....C.... .......... ........C. .A........ C.....C... .......... ....G..... ...T..C...
AB003288  C....C..G. .......... .........A .........C .....C.... .......... ........C. .......C.. C.....C... .......... ....G..... ......C...
AB003293  .....C..G. .......... .......... .........C .....C.... .......... ........C. .......C.. C.....C... .........G ....G..A.. ..........
AF006500  .....C..G. .......... .....A..A. .....A...C .....C.... .......... ........C. .......... C..T..... .......... .......A.. A.....C...
AB003290  C....C..G. .......T.. .......... .........C .A...C.... G......... ........C. .A.....C.. C.....C... .......... ....G..... ...T..C...
HGU94695  .....C..G. .......... .......... .....A...C .....C.... .......... ........T. .......C.. C.....C... .......... ....G..... ......C...
AB018667  ........G. .T........ .......... .....T...C ....CC.... .......... ........C. .A.....C.. C..T..C... .......... ....G..A.. ......AT..
AB021287  C.......T. .......... .......... .....C...C .....C.... .......... ........C. .A.....C.. .......... ....G..... .......A.. ......AT..
AB003292  C....T.... .......... .......... .....C...C .....C.C.. .......... ........C. .A.....C.. C........A .......... .......... ..........
AF070476  C.T.....G. ........AA A..CTC.... .......... .....C.... .......... ..C..T..T. .T..A..AAG G..A.....A ........A. ....G..T.. C..T......

          8800                                                                                                                 8919
          ϑ                                                                                                                    ϑ
HGU36380  TCCCCCCCCT ACATGGGGGT GGTTCATCAA TTGGATTTCA CAGCSCAGCG GAGTCGCTGG ---CGGTGGT TGGGGTTCTT AGCCCTGCTC ATCGTAGCGC TCTTTGGGTG AACTAAATTC
AB013500  .......... .......... .......... .......... ....C..... .......... ---....... .......... .......... .....G.... .......... ..........
AB003291  .......... .T........ ...A...... .......... ....C..... .......... ---....... .......... .......... .....G.... .......... ..????????
AB013501  .......... .......... .......... .......... ....C..... ......T... ---....... .......... .......... .....G.... .......... ..........
AF031829  .......... .......... .......... ........T. ..AGC...A. .......... ---....... .......... .......... ........C. ....C..... ..........
D90600    .......... .T........ .......... .......... ..AGC...A. .......... ---....... .......... .......... .....G..T. ....C..... ..........
D87255    .......... .T........ .......... .......... ..AGC...A. .......... ---......C .......... .......... ........C. ....C..... ..........
AF104403  .....T.... .T.....A.. .......... .......... ..AGC...A. .......... ---....... .......... .......... .....G..C. .T..C..... ..........
HGU63715  .......... .......... ...C...... .......... ..AGC...A. .......... ---....... .......... .......... ........C. ....C..... ..........
HGU44402  .......... .T........ ...A...... .......... ..AGC...A. .......... ---....... .......... .......... ........C. ....C..... ..........
AB003289  .......... .......... .......... .......... ..AGC..AA. .......... ---....... .......... .......... ........C. ....C..... ..????????
HGU45966  .......... .......... ...G...... ........T. ..AGC...A. .......... ---......C .......... .......... ........C. ....C..... ..........
D90601    .......... .T........ .......... .......... ....C..... .......... ---....... .......... .......... .....G.... .......... ..........
D87263    .......... .T........ .......... .......... ....C..... .......... ---....... .......... .......... .......... .......... ..........
AB008342  .......... .T........ .......... .......... ....C..... .......... ---....... .......... .......... .....G.... .......... ..........
AB003288  .....T.... .......A.. .......... .......... ....C..... ......T... ---....... .......... .......... .......... .......... ..????????
AB003293  .....T.... .......A.. ...C...... .......... ....C...A. A......... ---....... .......... T......... ........T. .......... ..????????
AF006500  .......... .T........ .......... ........T. ....C..... ......T... ---T...... .......... .A........ .......... .......... ..........
AB003290  .......... .......... .....C.... ........T. ....C..... .......... ---....... .......... .A........ .....G..A. .......... ..????????
HGU94695  .......... .......... .......... .......... ....C..... .......... ---....... .......... .......... .......... .......... ..........
AB018667  .....T.... .......A.. ...A..C... .......... ....C..... .......... ---....... .......... .A........ ........A. .......... ..........
AB021287  .......... .T........ ...A..C... .......... ....C..... .......... ---....... .......... ....A..... .....G.... ....C..... ..........
AB003292  .......... .......... ...A...... .......... ....C..... .......... ---....... .......... .......... .......... .......... ..????????
AF070476  G.T....... .T........ A..C..CAC. .......... .TCAG.CAA. ......G... CGA.T..TA. .T...C.A.. ......TG.A G.G..T..CA .......T.. .G.....-A.
```


8920
9039
HGU36380 ATCTGTTGCG GCAAGGTTGA GCGGCTGATC ACCGCTCAAG GAGGTTCCCG CCCTCCCCGC CCCAGGGGTC TCCCCGCTGG GTAAAAAGGG CCCGGCCTTG GGAGGCATGG TGGTTAC---
AB013500 .... .T..... ...A..... ....---
AB003291 ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
AB013501 ....C.G C..... ....GCTG.. ....---
AF031829 ....CCG .T.A..... .T.A..GG.. ....---
D90600 ....CCG .T.A..... .T.A..GG.. ....---
D87255 ....CTG .T.A..... .T.A.CGG.. ....---
AF104403 .....- --...CCGA. TT????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
HGU63715 ....CCG .T.A..... .T.A.CGG.. ....---
HGU44402 ....CTG .T.A..... .T.A.CGG.. ....---
AB003289 ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
HGU45966 ....CC. .T.A..... .T.A..GG.. ....???? ?????????? ?????????? ??????????
D90601 ....C.G CT.A..... .T.AGCTG.. ..... .T..... ....---
D87263 ....C.G C..A..... .T..GCTG.. ....---
AB008342 ....C.G C.T.A.C... ..T.GAGG.. .---..... ....---
AB003288 ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
AB003293 ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
AF006500 ....???????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
AB003290 ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
HGU94695 ....C.G CTA..... ..TAGCTG.. ....---
AB018667 ....C.G ...A..... .T...CTG.. ....---
AB021287 ....CCG .T.A..... .T.A.CGG.. ....--.. ....T..... ....---
AB003292 ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
AF070476 TCAG..ACG. ......AGCG CAA...---- ---------. .G.C..T... A..C....C. ..AG.AC.AT ....T. ....A....A.. ...TCG.C.A .T..GGAATC
9040
9159
HGU36380 ------TAAC CCCCTGGCAG GGTTAACGCC TGATGGTGCT AATGCACTGC CGCTTCGGCG GCGGGTCGCT ACCTTATAGC GTAATCCGTG ACTACGGGCT GCTCGCAGAG CCCTCCCCGG
AB013500 ------.... ....C.TA... ....C ....
AB003291 ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
AB013501 ------.... ....C..A... ....-....-.. ....
AF031829 ------.... ....C..A... .... .A..A...T. ....
D90600 ------.... ....C..A... .... .A..A...T. ....
D87255 ------.... ....C..A... ...C..... ..... .A.....T. ....A. ....A ....
AF104403 ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
HGU63715 ------.... ....C.T.... .... .A..A...T. ....
HGU44402 ------.... ....C..A... .... .A.....T. ....
AB003289 ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
HGU45966 ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
D90601 ------.... ....C..A... ....--.TC... ....
D87263 ------.... ....C..A... ....--.TC... .... C..... G.....
AB008342 ------.... ....A.T.... ....--.TC... ....
AB003288 ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
AB003293 ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
AF006500 ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
AB003290 ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
HGU94695 ------.... ....A... ....?? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
AB018667 ------.... ....C..A... .... .A.....T. ....A. A........A ....
AB021287 ------.... ....C..A... ....A..... ..... .A.....T. .....T.. ....G. .....T.. A.....
AB003292 ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
AF070476 GTAAAA.CCA ....TA... ....G. ...C...C.. T..--..CA. ..C..... ..G..... ....T...

```
          9160                                                                                                  9234
          ϑ                                                                                                     ϑ
HGU36380  ATGGGGCACA GTGCACTGTG ATCTGAAGGG GTGCACCCCG GTAAGAGCTC GGCCCAAAGG CCGGGTTCTA CT???
AB013500  .......... .......... .......... .......... .......... .......... .......... ..???
AB003291  ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????
AB013501  .......... .......... .......... .......... .......... .......... .......... .????
AF031829  .......... .......... .......... .......... .......... .????????? ?????????? ?????
D90600    .......... .......... ?????????? ?????????? ?????????? ?????????? ?????????? ?????
D87255    .......... .......... .......... .........? ?????????? ?????????? ?????????? ?????
AF104403  ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????
HGU63715  .......... .......... .......... .......... .......... .....G.... .......... .????
HGU44402  .......... .......... .......... .......... .G........ .....G.... ....S..... .????
AB003289  ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????
HGU45966  ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????
D90601    .......... .......... .......... .......... .......... .......... .......... ..???
D87263    .......... .......... .......... .......... .......... .......... .......... ..???
AB008342  .......... .......... .......... .......... .......... .......... .......... ..???
AB003288  ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????
AB003293  ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????
AF006500  ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????
AB003290  ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????
HGU94695  ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????
AB018667  .......... .......... .......... .......... .......A... .......... .......... ..???
AB021287  ......G... C......... .......A.. ...T...... .......A.. T......... .......... ..???
AB003292  ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????
AF070476  ...A...G.. .......??? ?????????? ?????????? ?????????? ?????????? ?????????? ?????
```

RNA (2002), *8*:824–841. Cambridge University Press. Printed in the USA.

DOI: 10.1017.S1355838202024068

Thermodynamic and phylogenetic prediction of RNA secondary structures in the coding region of hepatitis C virus

ANDREW TUPLIN,[1] JONNY WOOD,[2] DAVID J. EVANS,[3] ARVIND H. PATEL,[2] and PETER SIMMONDS[1]

[1]Laboratory for Clinical and Molecular Virology, University of Edinburgh, Summerhall, Edinburgh, EH9 1QH, Scotland
[2]MRC Virology Unit, University of Glasgow, Church Street, Glasgow, G11 5JR, Scotland
[3]Department of Virology, University of Glasgow, Church Street, Glasgow, G11 5JR, Scotland

ABSTRACT

The existence and functional importance of RNA secondary structure in the replication of positive-stranded RNA viruses is increasingly recognized. We applied several computational methods to detect RNA secondary structure in the coding region of hepatitis C virus (HCV), including thermodynamic prediction, calculation of free energy on folding, and a newly developed method to scan sequences for covariant sites and associated secondary structures using a parsimony-based algorithm. Each of the prediction methods provided evidence for complex RNA folding in the core- and NS5B-encoding regions of the genome. The positioning of covariant sites and associated predicted stem-loop structures coincided with thermodynamic predictions of RNA base pairing, and localized precisely in parts of the genome with marked suppression of variability at synonymous sites. Combined, there was evidence for a total of six evolutionarily conserved stem-loop structures in the NS5B-encoding region and two in the core gene. The virus most closely related to HCV, GB virus-B (GBV-B) also showed evidence for similar internal base pairing in its coding region, although predictions of secondary structures were limited by the absence of comparative sequence data for this virus. While the role(s) of stem-loops in the coding region of HCV and GBV-B are currently unknown, the structure predictions in this study could provide the starting point for functional investigations using recently developed self-replicating clones of HCV.

Keywords: coding; covariant; GBV-B; stem-loop; synonymous; thermodynamic

INTRODUCTION

Infection with hepatitis C virus (HCV) has been identified as the principal cause of posttransfusion non-A, non-B hepatitis (Choo et al., 1989; Kuo et al., 1989). It is also a major cause of chronic hepatitis, cirrhosis, and hepatocellular carcinoma throughout the world. HCV has been classified as a member of the *flaviviridae*, with a plus-sense RNA genome containing a single open reading frame (ORF). Details of the replication of HCV are currently poorly understood, principally because of the lack of a method for its in vitro culture. Recently a subgenomic RNA of HCV genotype 1b lacking the region encoding the core, E1, and E2 proteins was shown to be capable of self-replication in the human hepatoma cell line, HuH-7 (Lohmann et al., 1999; Blight et al., 2000), although without structural proteins it lacked the ability to produce infectious virus.

The genome of HCV forms RNA secondary structures in the 5′ and 3′ untranslated regions (UTRs) that are likely to play a role in initiation of RNA replication, and in the 5′ UTR, for ribosomal binding associated with its IRES function (Tsukiyama Kohara et al., 1992; Reynolds et al., 1996). Unusually, efficient functioning of the HCV IRES is dependent on sequences in the downstream coding sequence (Honda et al., 1996a; Reynolds et al., 1996), suggesting that RNA secondary structures in the core gene may contribute to IRES structure. Little attention has hitherto been paid to the existence and functional importance of RNA secondary structure in other parts of the genome. Through analysis of variability at synonymous sites, and by analysis of covariance to determine sites of internal base pairing, we obtained evidence for extensive RNA secondary structure formation in the coding region of a flavivirus related to HCV, described as hepatitis G virus

Reprint requests to: Peter Simmonds, Laboratory for Clinical and Molecular Virology, University of Edinburgh, Summerhall, Edinburgh, EH9 1QH, Scotland; e-mail: Peter.Simmonds@ed.ac.uk.

(HGV) or GB virus-C (GBV-C; Simmonds & Smith, 1999; Cuceanu et al., 2001). The essential role that a relatively small stem-loop in the middle of the coding region of poliovirus (Goodfellow et al., 2000) plays in its replication (Rieder et al., 2000) suggests that the structures found in HGV/GBV-C may have equally significant roles in its life cycle. It is similarly possible that RNA folding may represent functional components of a much wider range of flaviviruses and other positive-stranded RNA viruses.

In this study, we have therefore applied a number of separate phylogenetic and thermodynamic prediction methods to investigate the extent to which the HCV genome may also be folded by internal base pairing. Predictions of secondary structure that we have made are amenable to future functional study through mutational analysis of replicating clones of HCV.

RESULTS

Suppression of synonymous variability

Reduction in the sequence diversity of synonymous sites may result from constraints on sequence change that arise from the formation of stem-loop structures that influence virus phenotype (Simmonds & Smith, 1999). HCV sequences of different genotypes are highly divergent in sequence, particularly at synonymous sites (Smith et al., 1997). We therefore developed a method to score synonymous variability that depends on the overall phylogeny of the HCV sequences and reconstruction of the nucleotide sequence of each codon at each ancestral node (Fig. 1). This method is more capable of detecting multiple substitutions at each site than simple pairwise comparison. It also correctly scores

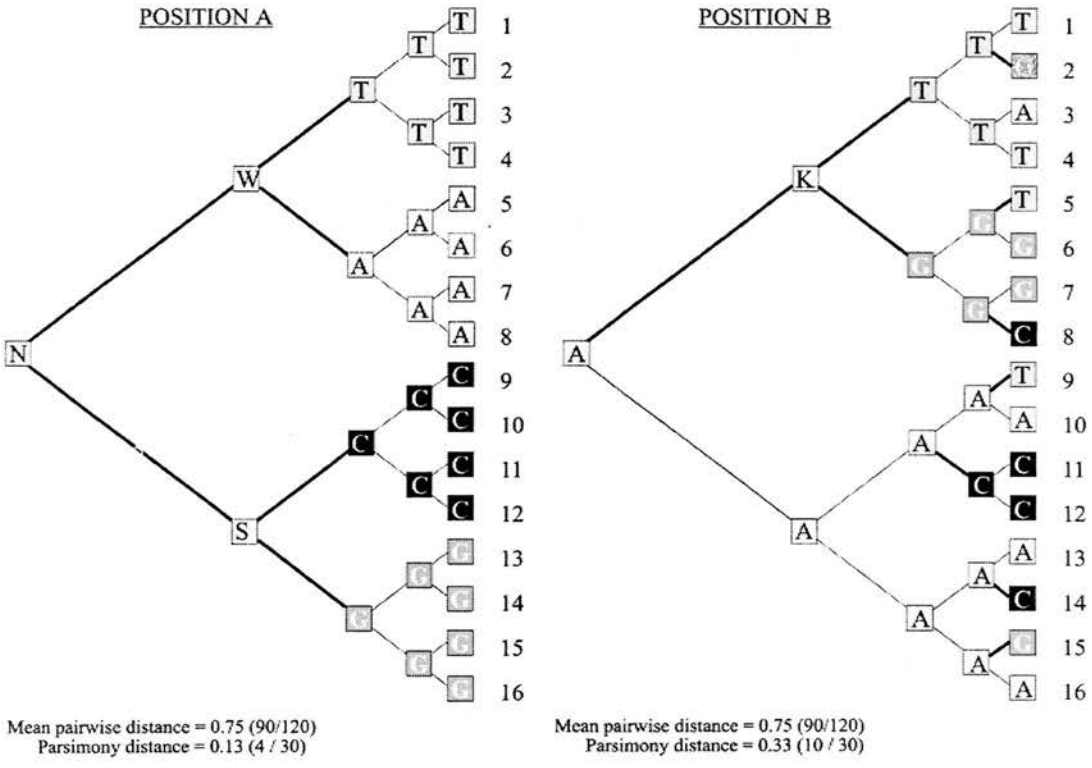


FIGURE 1. Examples of the calculation of parsimony distances through reconstruction of sequence substitutions at two separate nucleotide positions A and B. For the 16 sequences analyzed (1–16), the overall phylogeny and sequences of ancestral nodes were estimated by a standard parsimony program (DNAPARS). At the two nucleotide positions shown (A, B), nucleotides in each sequence and those reconstructed for each immediately ancestral node were compared to estimate minimum number of sequence changes to produce observed pattern of sequence diversity. Position A: Nucleotide site where variability is congruent with overall phylogeny and where 4 nucleotide changes can be reconstructed to produce a parsimony distance of 0.13 (4 nucleotide changes in 30 comparisons). Position B: Site where variability is noncongruent with phylogeny and where a minimum of 10 nucleotide changes are required (parsimony distance 0.33). By contrast, conventional measurement of pairwise distances at the both sites A and B produce values of 0.75, representing saturation. The measurement of pairwise distances at nucleotide position A is an overestimate because it treated every difference between sequences 1–16 as phylogenetically independent.

individual substitutions that may occur deep in the phylogeny of a particular clade, and therefore avoids the multiple scoring associated with averaging matrices of pairwise distances.

Synonymous variability between HCV sequences of different genotypes showed considerable differences across the coding region of the HCV genome (Fig. 2A). Consistent with the above analysis, heterogeneity in synonymous variability was more apparent using parsimony than equivalent analyzes using pairwise distances (data not shown). Suppression of synonymous variability relative to the rest of the genome was observed between nt 1 and 530, with particularly marked dips at the start of the coding sequence and at position 364. In the E1-, E2-, NS2-, and NS3-encoding regions, mean variability over 50 codon windows ranged from 0.27 to 0.34. Beyond position 4900, there was a trend for a consistent reduction in mean synonymous variability with particularly marked dips at the 5′ and 3′ ends of the NS5B-encoding region (7768 and 9091). The degree of suppression of synonymous variability at the end of the NS5B-encoding region was comparable to that observed in the core gene (Fig. 2A).

The availability of a large number of epidemiologically unlinked complete genome sequences of type 1b allowed a separate analysis of intrasubtype variability (Fig. 2B). HCV genotype 1b sequences are thought to have arisen from a common ancestor approximately 60–70 years ago (Smith et al., 1997), a time span over which covariant substitutions are unlikely to have accumulated (Simmonds & Smith, 1999). This comparison may therefore provide a more sensitive indication

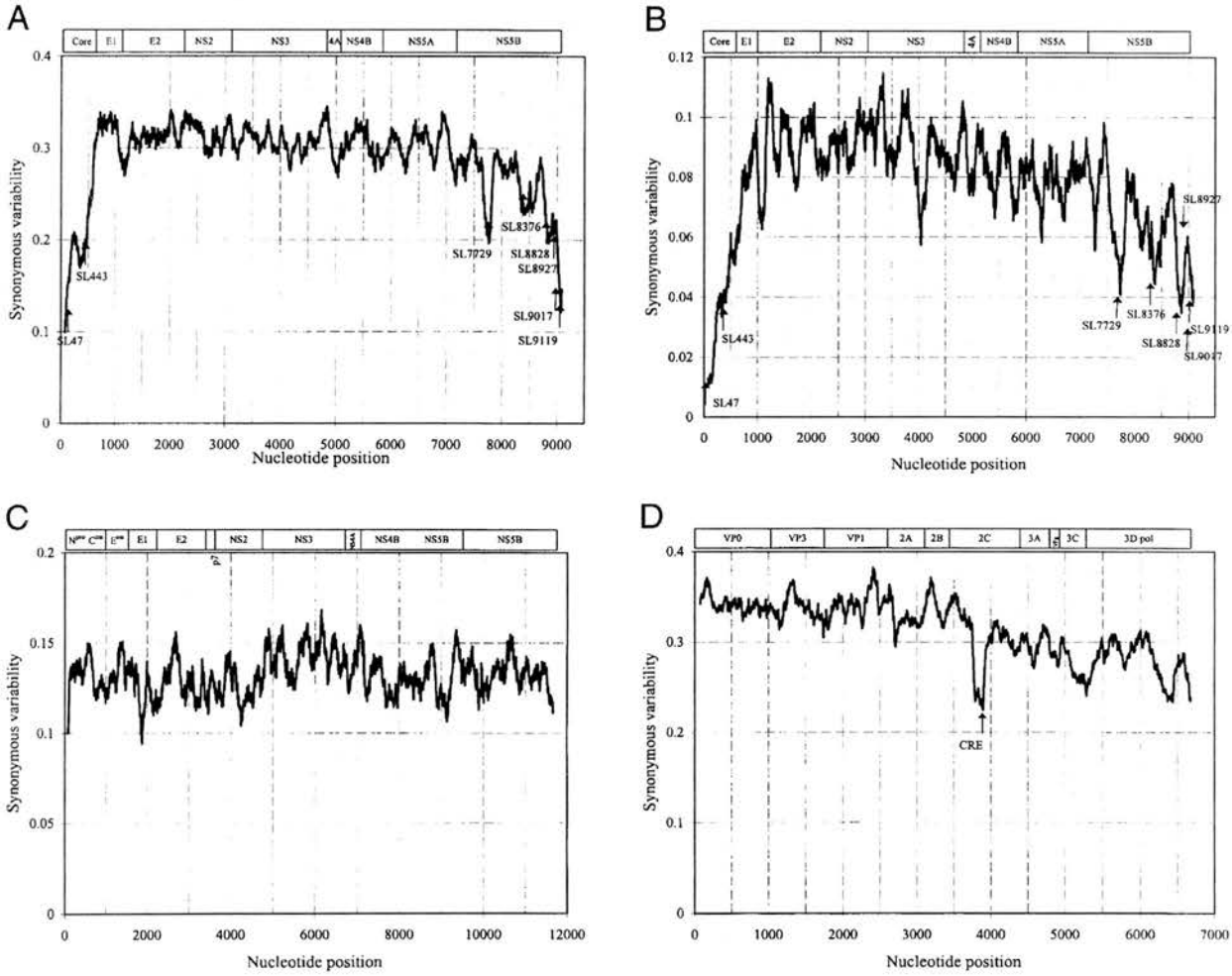


FIGURE 2. Variability at synonymous sites estimated by parsimony between coding regions of (**A**) single examples of HCV genotypes and subtypes; location of covariant sites (Fig. 5) indicated by arrows. The approximate sites for cleavage of structural (core, E1, E2) and nonstructural proteins (NS2-NS5B) indicated in upper panel (**B**) 65 epidemiologically unlinked genotype 1b sequences, (**C**) pestivirus genotypes 1–4, and (**D**) human enteroviruses (poliovirus, Coxsackieviruses A and B, echoviruses).

of the regions where sequence change is constrained. There was an overall similarity with the previous comparison between HCV genotypes in the regions where synonymous variability was suppressed (compare Fig. 2A with 2B). However, there was much greater variability between different parts of the coding region. For example, the degree of synonymous variability at the start of the core gene between type 1b sequences was greater than 10-fold lower than the mean values for E1 and E2, a larger differential than observed between HCV genotypes (approximately 3.5-fold; Fig. 2A). Much more pronounced dips in variability were observed elsewhere in the genome, not only in the NS5B-encoding region but also in NS3- and NS4A/B-encoding regions (Fig. 2B). To investigate whether regions where variability was suppressed were shared between all type 1b sequences, the set of 65 sequences were randomly assigned to three separate groups and synonymous variability was independently measured in each group. There was a strong correlation between synonymous variability in the three independent samplings of type 1b sequences, and an absence of any sites with discrepant synonymous variability values (data not shown). These observations provide evidence for similar sequence constraints on type 1b sequences in each of the three samplings.

To investigate whether the observed suppression in synonymous variability at the ends of the HCV genome resulted from biased codon usage in these regions, base composition at third codon positions was calculated for HCV genotypes 1–6 (Fig. 3A). Base composition (including G + C and purine content) was similar over the length of the genome, with no evidence for different codon usage in regions where synonymous variability was suppressed compared with more variable regions. Similarly, there was little variability in the base composition at first and second codon positions in different parts of the HCV genome (data not shown).

Suppression of synonymous variability might also originate from biased dinucleotide frequencies that limit codon choice. However, 14 of the 16 dinucleotides showed little or no differences from their expected frequencies calculated from their base composition at each of the three codon positions (i.e., dinucleotides at first and second codon positions, at second plus third, and third plus one; note only the latter two are subject to selection independently of coding capacity; Fig. 4). The CG dinucleotide was underrepresented at all three codon positions (0.67, 0.66, and 0.73; mean 0.72), whereas UG was slightly overrepresented (1.31, 1.18, and 1.21; mean 1.23). However, there was no correlation between the observed differences in synonymous variability in the HCV genome with frequencies of the CG and UG dinucleotides (Fig. 3B), nor with any of the other 14 dinucleotides (data not shown).

Analysis of the distribution of synonymous variability was extended to sequences from pestiviruses (Fig. 2C) and enteroviruses (Fig. 2D). In contrast to HCV, there was no suppression of variability at the ends of the pestivirus genomes, nor was there evidence for more restricted regions of suppression in the genes corresponding to NS5A or the 5′ end of NS5B (Fig. 2C). Enterovirus sequences showed great variability at synonymous sites (Fig. 2D), and similarities in the pattern of diversity with HCV. In particular there was an overall reduction in variability in the 3′ end of the genome, with marked areas of suppression at positions 3841, 5339, 6467, and the extreme 3′ terminus of the genome. The dip at position 3841 occurred in a region of RNA secondary structure (Goodfellow et al., 2000) with a role in the initiation of RNA transcription (Rieder et al., 2000). Enterovirus sequences, however, did not show the extreme suppression of synonymous variability at the 3′ terminus, and showed no reduction in variability at the 5′ end (corresponding to the start of the genome coding for the nucleocapsid).

Thermodynamic prediction of RNA secondary structure

The existence of RNA secondary structure in the coding region of HCV was independently investigated by comparison of free energy on folding of overlapping 500-base-coding-sequence fragments with those of sequence-order-randomized controls. In this study, we have developed a number of methods to randomize the sequence order of the HCV coding sequence to determine its contribution to RNA folding. Many of these were developed to prevent the randomization process altering other sequence attributes, such as regional differences in base composition and biased dinucleotide frequencies, that may have a compounding effect on free energy calculation. The methods used were as follows:

1. *Nucleotide order randomization (NOR):* Randomization of nucleotide sequence order. This is the standard method used in most previous studies.
2. *Codon order randomization (COS):* Randomization of codon order, therefore avoiding disruption of sequence order within triplets.
3. *Like-codon randomization (CLR):* Randomization of the order of codons specifying each amino acid. Following randomization, the encoded amino acid sequence remains unaltered.
4. *Like-codon swap (CLS):* Pairwise exchange of like codons (e.g., the first glycine codon in the sequence is exchanged with the second, the third with the fourth, etc.). Alternatively, the second is exchanged with the third, the fourth with the fifth, etc.).
5. *Dinucleotide randomization (CDR):* Randomization of each set of codons with identical first and third bases (i.e., the 16 sets with the sequences AnA, AnT, AnG, ... TnG, TnT). The randomized sequence

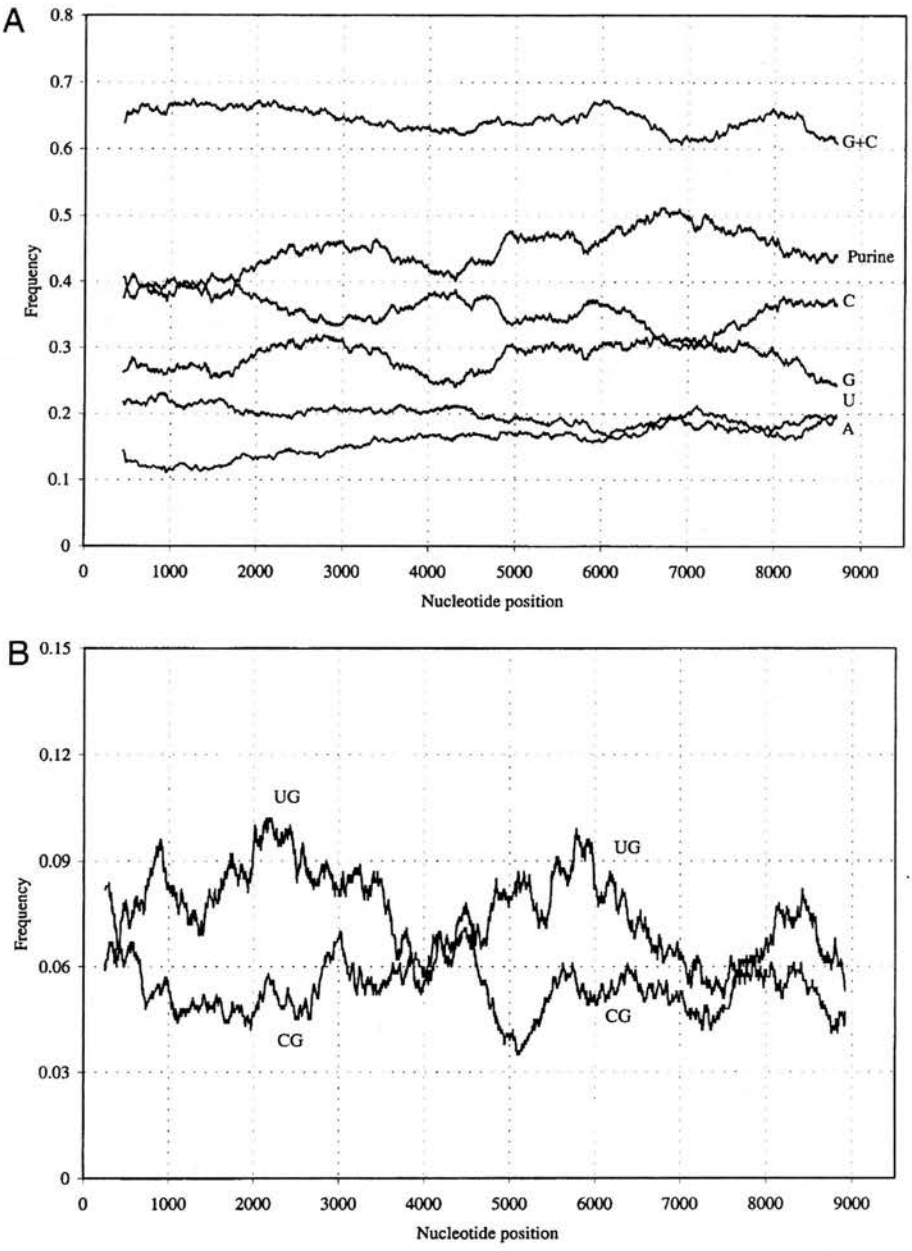


FIGURE 3. A: Scan of base composition at third codon positions in coding sequences of HCV genotypes 1–6 (mean values shown), including combined values for purine (G + A) and G + C. **B:** Scan of CG and UG dinucleotide frequencies across HCV genome (mean of genotypes 1–6, all three codon position).

will have an identical dinucleotide composition (but different encoded amino acid sequence) from the native sequence.

6. *Dinucleotide swap (CDS):* Pairwise exchange (as CLS) of each of the 16 sets of codons with identical first and third bases.

Methods NOR, COS, CLR, and CDR can be applied multiple times to a native sequence. The difference in free energy of folding between the native sequence and each randomized sequence approximates to a normal distribution (Rivas & Eddy, 2000), providing an empirical statistical test for the significance of the observed differences. Accordingly, differences in free energy between native and randomized sequences can be expressed as a *Z*-score (Workman & Krogh, 1999), which is the number of standard deviations by which the predicted free energy of the native sequence is lower than

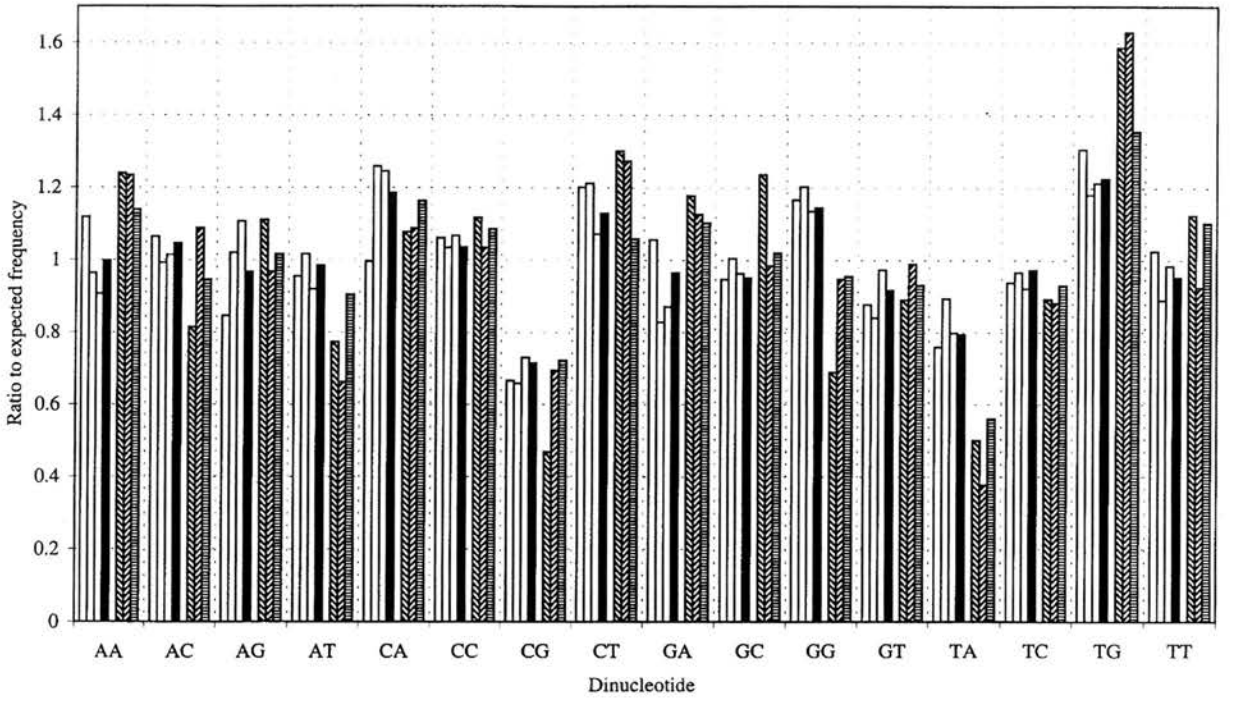


FIGURE 4. Dinucleotide frequencies of HCV and three mammalian genes (albumin, alphaglobin and actin), expressed as ratios to expected value calculated from base composition. For each dinucleotide: columns 1–3 (unfilled): dinucleotide ratios of HCV for each codon position (first and second base position, second/third, and third/first); column 4 (filled): mean for all three positions; columns 5–7: mean values for all three codon positions for coding sequences of albumin (descending diagonal stripes), alphaglobin (ascending diagonal stripes), and actin (horizontal stripes).

the mean of the randomized sequences. This measure therefore takes into the account the differences in free energy as well as the range of values between independent randomizations. The randomization methods that exchange like-coding or like-dinucleotide codons generate only two randomized sequences, preventing the calculation of a *Z*-score.

Applying the methods CLR and CLS to HCV coding sequences retains their encoded amino acid sequences, and in contrast to the NOR method, retains the base composition at first, second, and third codon positions. Method CLS minimizes the distance over which codons can be exchanged to pairwise swaps and therefore has the advantage of retaining any nonhomogeneities in base composition that may exist in sequence. Methods CDR and CDS are similarly codon orientated, retain dinucleotide composition, codon composition, but not codon order, and, in the case of CDS, also preserve local differences in base composition.

To investigate the relationship between nucleotide sequence order and folding free energy, we used each of the six methods to randomize sequences with no known or likely RNA secondary structure (the coding sequences in the mammalian albumin, actin, HLA class II, and alphaglobin genes), and parts of the coding region of HGV/GBV-C (5′ and 3′ ends with previously demonstrated RNA secondary structure; Simmonds & Smith, 1999; Cuceanu et al., 2001; Fig. 5A, B). Each of the six sequence order randomization methods produced comparable differences in folding free energy from native sequences. In the case of mammalian sequences, none of the methods produced a positive differences in free energy of greater than 3% in any of the sequences (Fig. 5A); similarly, *Z*-scores were limited to values of greater than -1 for each of four methods where this statistic could be calculated (NOR, COR, CLR, and CDR), indicating no significant difference in folding free energy between native and randomized sequences (Fig. 5B). In contrast, large differences in folding free energy were observed between native HGV/GBV-C 5′ and 3′ sequences and those randomized by each of the six methods (10–17%); *Z*-scores ranging from -3.7 to -5.1 indicated that each of these differences was statistically significant ($p < 0.01$; Workman & Krogh, 1999; Rivas & Eddy, 2000). Each of the six methods also provided evidence for sequence-order-dependent secondary structure in the corresponding positions in the HCV genome (free energy differences 10–15%, *Z*-scores 3.4–6.3).

Because of the close concordance of free energy differences (and *Z*-scores where calculable) between the different scrambling methods, it is unlikely that un-

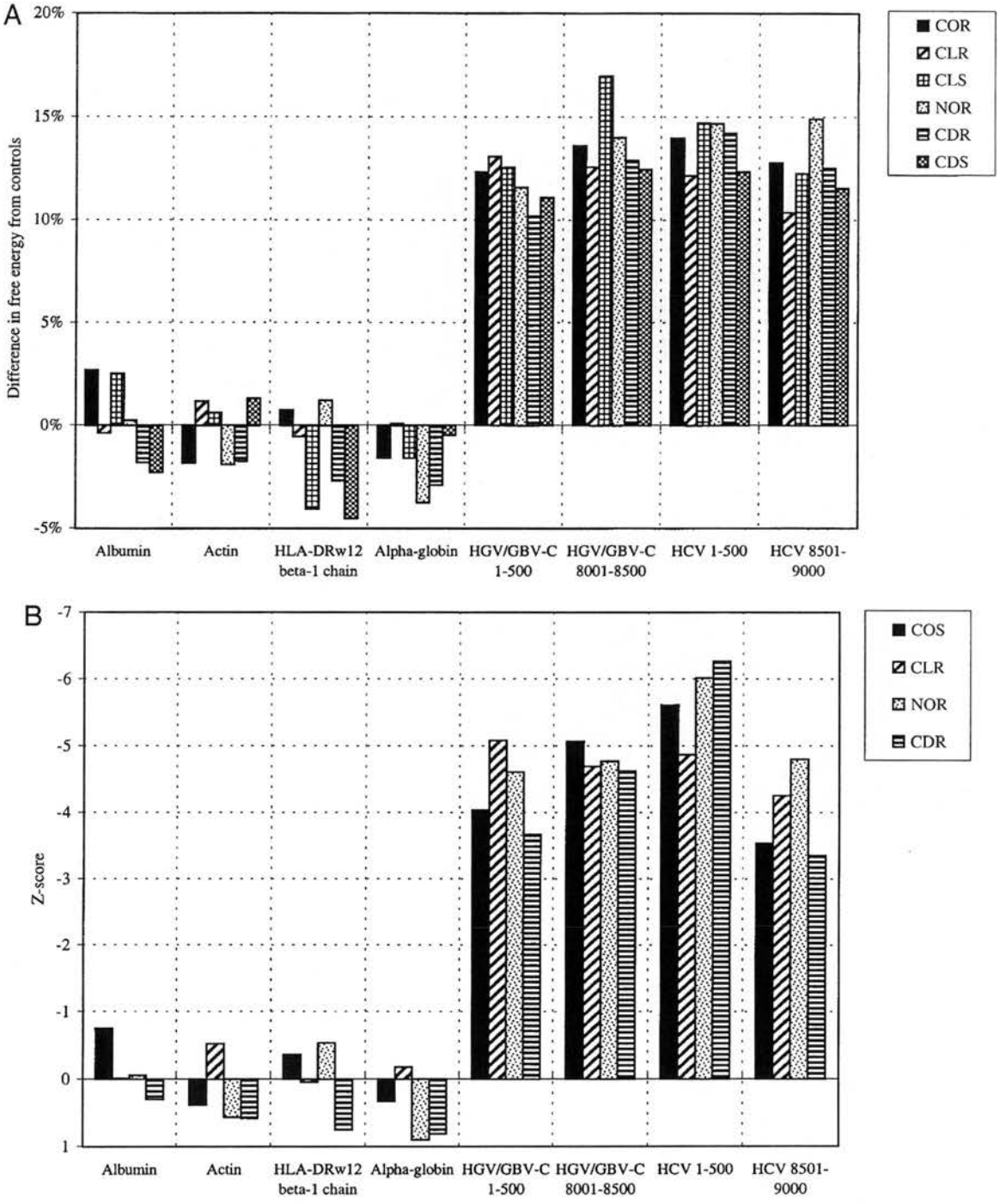


FIGURE 5. A: Differences in folding free energy between native mammalian (albumin, actin, HLA class II, alphaglobin) and viral coding sequences (5′ and 3′ ends of HGV/GBV-C and HCV) and those randomized by six different scrambling methods (NOR, COR, CLR, CLS, CDR, CDS). **B:** Corresponding *Z*-scores for NOR, COR, CLR, and CDR.

wanted effects of sequence order randomization, such as disruption of codon composition, dinucleotide frequencies, or regional differences in base composition, accounted for the observed excess of free energy in native viral sequences. HCV actually lacks the extreme composition biases that are found in mammalian sequences used as controls. The mean G + C content of HCV ranged from 0.50 to 0.62 at the three codon po-

sitions, well within the range of G + C compositions of albumin (mean 45.4%; 34.2% at third base positions) and alphaglobin (mean 62.0%; 83.0% at third base positions). Additionally, the self-complementary dinucleotides (GC, CG, AU, and UA) that potentially increase free energy on folding were not overrepresented in HCV (Fig. 4); indeed, the CG dinucleotide occurred at a lower than expected frequency (mean 0.72). As with base composition, biases in HCV dinucleotide frequencies were less marked than those found in mammalian genes. For example, coding sequences of human alphaglobin, actin, and albumin showed similar or greater underrepresentation of CG, much greater suppression of UA dinucleotides (0.38–0.56), and there was much greater overrepresentation of UG (1.34–1.63).

For more detailed analysis of RNA structure in the HCV genome, we used the two least disruptive randomization methods that allowed *Z*-scores to be calculated (CLR and CDR). Coding sequences of HCV genotypes 1a, 2a, 3a, 4a, 5a, and 6a were divided into 500-base fragments, overlapping by 250 bases (36 fragments over an alignment length of 9,168 nt). Folding free energies were compared with those of 50 replicates of each fragment randomized in sequence using the CLR and CDR methods (Fig. 6). Each of the six genotypes showed between 6.2 and 8.8% difference in folding free energy between native and scrambled sequences over the length of the genome (mean values: CLR: 7.8%, CDR: 6.9%; mean *Z*-scores: CLR: −2.58; CDR: −2.32). To localize potential secondary structure, mean values of each of the six genotypes were plotted against genome position (Fig. 7). The greatest differences in free energy were observed at the extreme 5′ end of the genome (fragments 1–500, 250–750) and at the 3′ end (fragments 7251–7750 and onwards), with good concordance between the two randomization methods. Free energy differences showed a close, inverse correlation with suppression of synonymous variability (Figs. 2A, B, and 7A).

Parallel testing of native sequences in reverse, complement orientation showed consistently lower differences on folding from (reverse complemented) sequence-order-randomized controls (Figs. 6, 7B). This reduction in folding energy difference was observed in all HCV genotypes using both randomization methods (Figs. 6, 7). It was also consistently observed throughout the coding region of the HCV genome, with values rarely exceeding 6%, and *Z*-scores invariably below −3, and generally below −2. These observations indicate that RNA structure is not only distributed through-

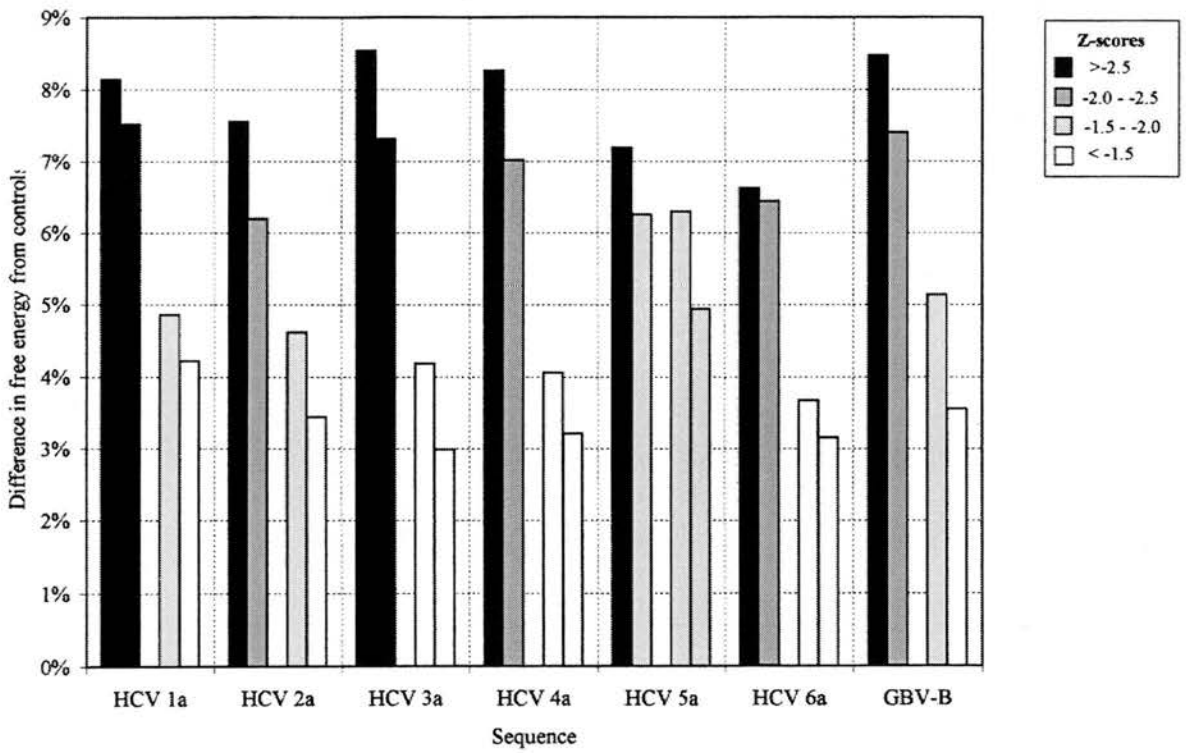


FIGURE 6. Mean difference in folding free energy of 500-base fragments spanning viral genome of HCV genotypes 1a–6a, and GBV-B, using two scrambling methods (CLR, CDR). For each sequence, columns 1, 2: native sequence; columns 3, 4: reverse complement sequence. *Z*-score ranges indicated by shading.

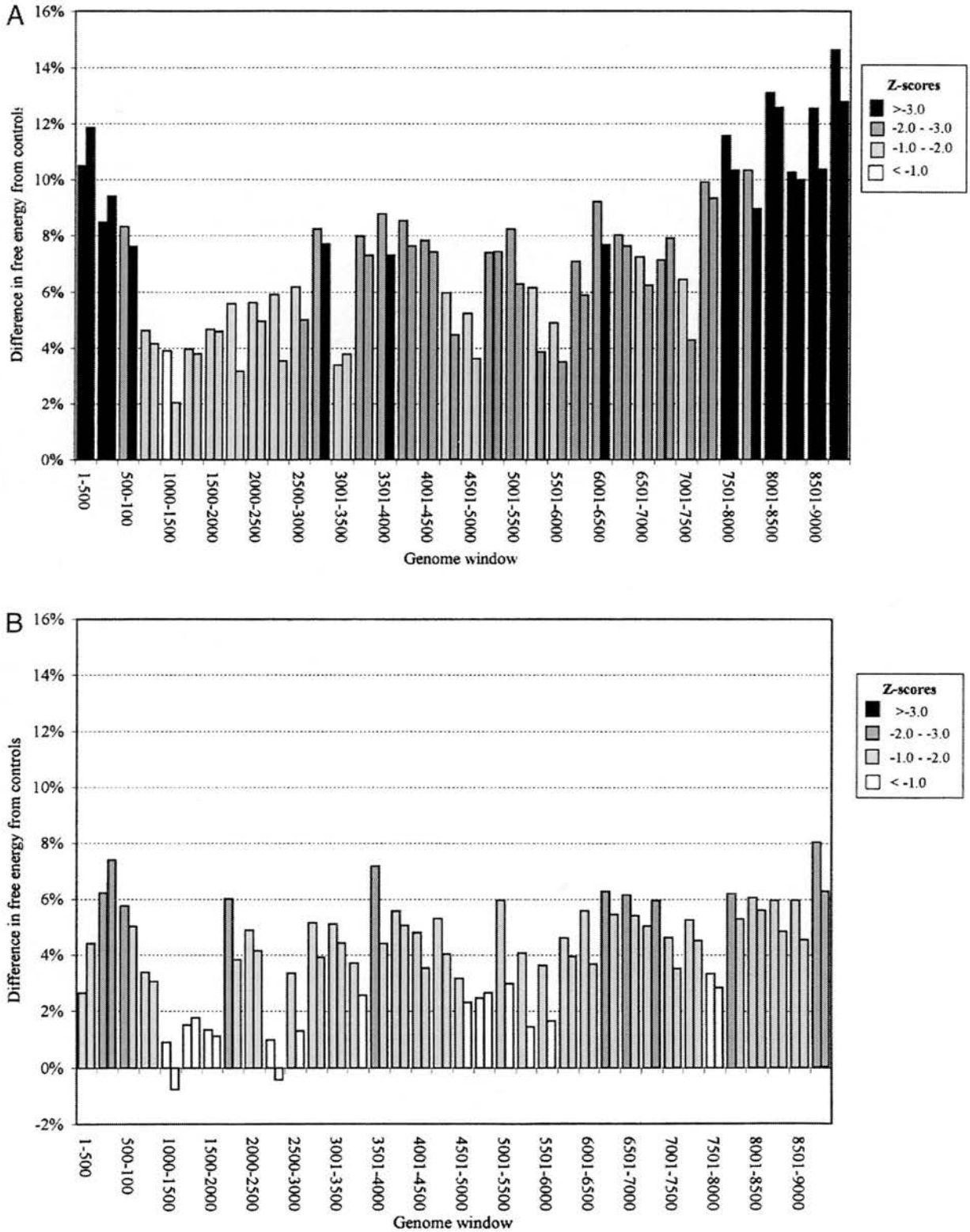


FIGURE 7. **A:** Mean difference in folding free energy of 500-base fragments of HCV genotypes 1–6 in different regions of the HCV genome using two scrambling methods [CLR (column 1), CDR (column 2)]. *Z*-score ranges indicated by shading. **B:** Free energy differences and *Z*-scores of corresponding reverse, complement sequences.

out the HCV genome, but is probably the most relevant biologically for HCV RNA in its positive sense orientation.

Covariance scanning

The existence of paired covariant sites associated with adjacent regions of potential base pairing provides independent evidence for the location of regions of secondary structure. The problem of phylogenetic structure and nonindependence of substitutions among members of different HCV clades previously encountered in the analysis of synonymous variability also presented difficulties with scoring covariant substitutions. Accordingly, covariant changes at paired base positions were only scored between each sequence or node and their immediate ancestors reconstructed by parsimony. As a result, the covariant score reflects the minimum number of evolutionary steps underlying the observed substitutions. Compared with HGV/GBV-C (Simmonds & Smith, 1999), fewer covariant sites were detected among the HCV sequences analyzed using equivalent input settings (Fig. 8). Using a variety of scanning parameters, a total of 14 covariant sites in eight potential stem-loop structures were detected in the coding region of the HCV genome, located in the core and NS5B regions. This compares with 48 sites in 23 potential stem-loops in the coding region of HGV/GBV-C (data not shown). The greater number of covariant sites in HGV/GBV-C may indicate more extensive secondary structure, or a lack of conservation of stem-loops between genotypes of HCV (see Discussion).

HCV stem-loop structures formed by covariant base pairings were located precisely in regions where synonymous variability was suppressed (Fig. 2A). Partic-

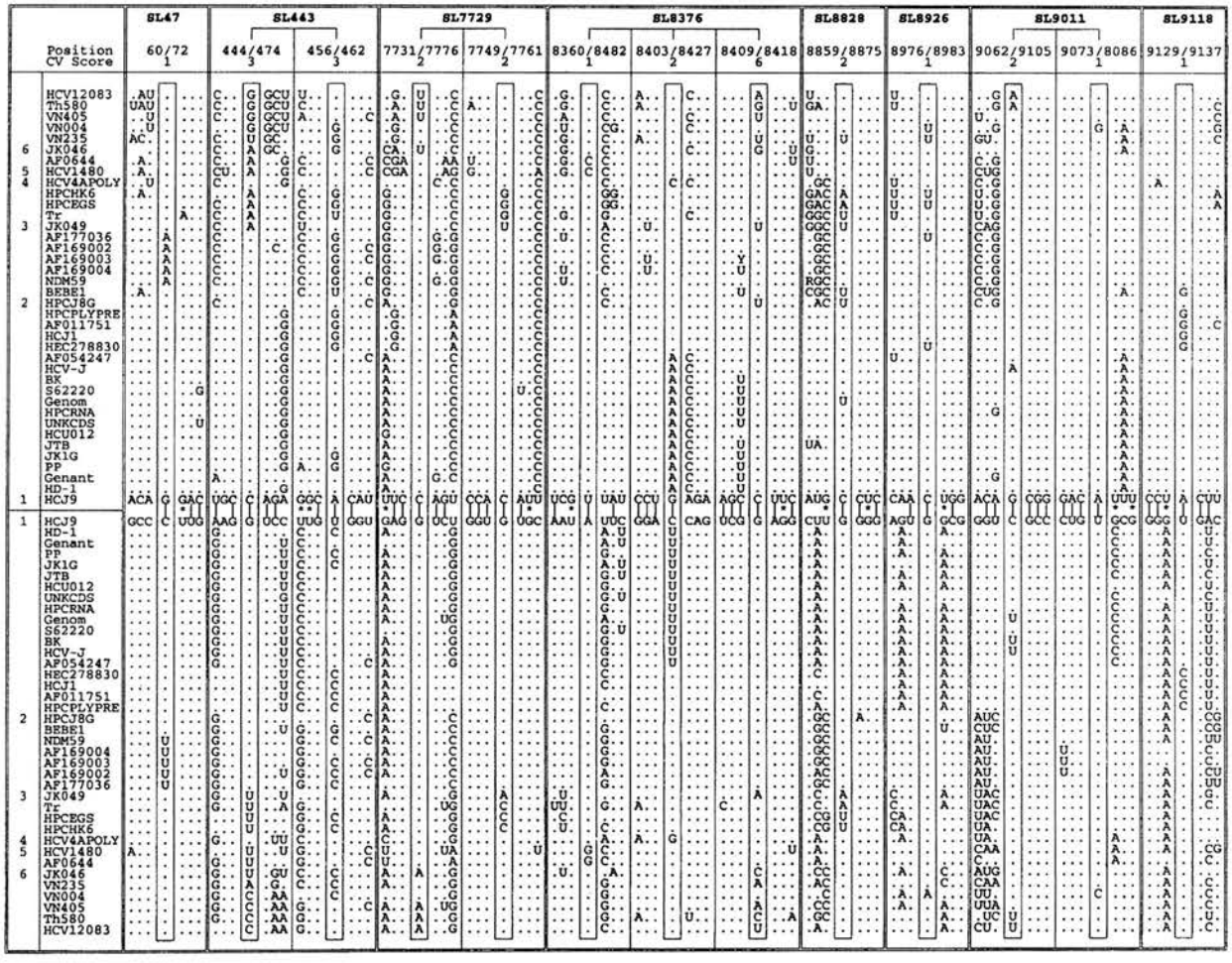


FIGURE 8. Covariant sites in HCV genome identified by parsimony. Stem-loop (SL) numbers and positions of upstream and downstream paired bases are shown above sequences. Covariant sites occurring in the same stem-loop are indicated by grouping into boxes. Genotype (1–6) of each sequence indicated on left. Covariant (CV) scores represent number of independent substitutions at covariant site (G ↔ C/U covariant changes not scored).

ularly remarkable was the concentration of covariant sites at the 3′ end of the NS5B gene that showed the most precipitous decline in synonymous variability.

Prediction of HCV RNA secondary structure

The combination of synonymous variability, differences in free energy, and identification of paired nucleotides by covariance scanning identified a number of discrete regions in the coding sequence of the HCV genome where secondary structure formation was likely to occur. Accordingly, thermodynamic secondary structure predictions of genome segments of genotypes 1–6 from the core- and NS5B-encoding regions were created using the program MFOLD using standard parameters (Fig. 9). Secondary structure predictions were made for stem-loop structures showing covariance (SL47, SL443, SL7730, SL8376, SL8828, SL8926, SL9011, and SL9118). Between genotypes, the structures varied in length, in the degree of sequence conservation

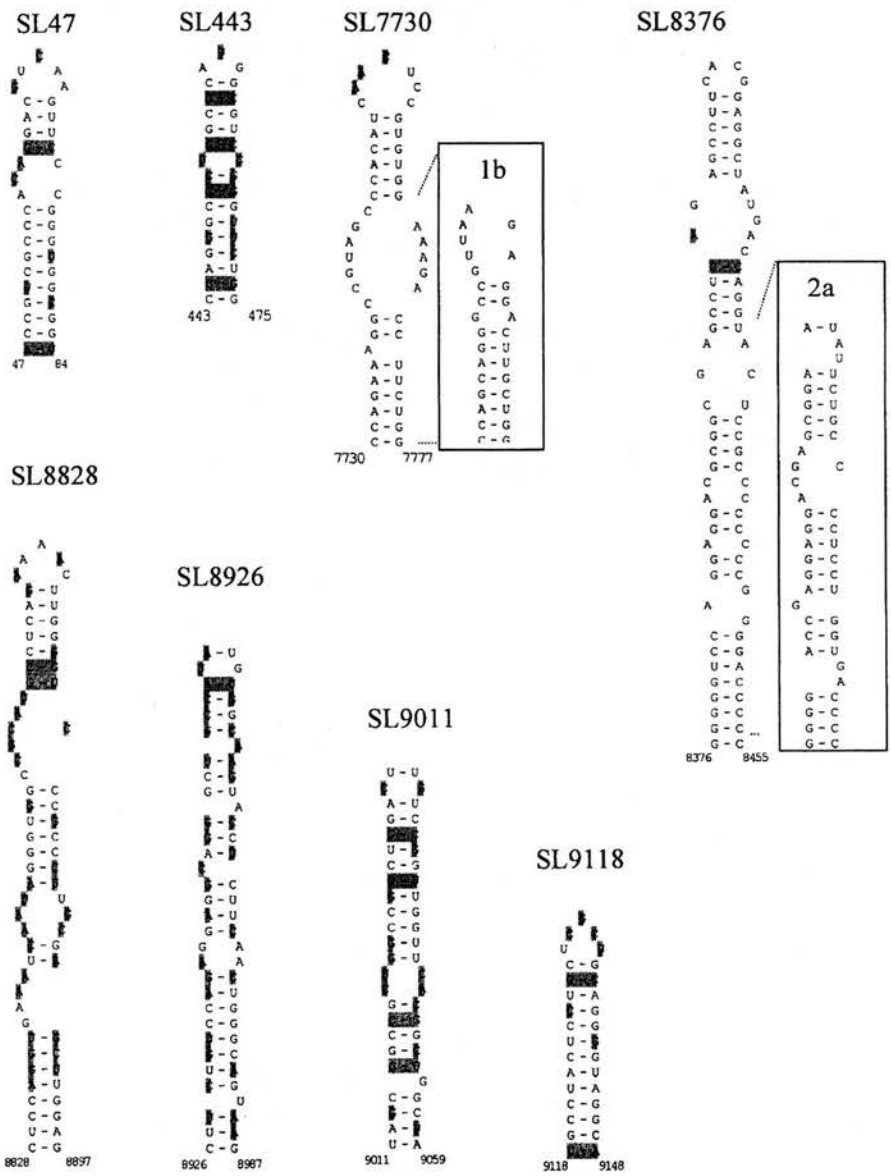


FIGURE 9. Predicted stem-loop (SL) structures conserved between HCV genotypes 1–6 containing the covariant sites shown in Figure 8. Stem-loop numbers and positions of upstream and downstream extents of base pairing shown above and below sequences. Covariant sites are indicated by shading of paired nucleotides. Example of structurally different stem-loops (SL7730 and SL8376) indicated in boxes.

of the unpaired regions, and in some cases, the predicted base pairings (Fig. 10). In the predicted base-pairing regions, third codon positions were most commonly aligned with downstream third codon positions, so that covariant changes were usually synonymous at both sites (in four from the seven predicted loops). However, nonsynonymous covariant substitutions were observed in SL8828, SL8926, and SL9118, resulting from base pairing between nucleotides at different codon positions.

Secondary structure prediction for GBV-B

Methods to analyze the secondary structure of the virus most closely related to HCV, GBV-B, are limited by the absence of comparative sequence data from independent isolates or genotypes of the virus (Simons et al., 1995). However it was possible to analyze the single complete genome sequence available thermodynamically (Fig. 6). The coding region of the GBV-B sequence showed differences in free energy from sequence-order-randomized controls similar to those observed for HCV, with comparable differences between sense and antisense sequences. Folding free energy differences showed a similar genome distribution to that of HCV, with the largest values and Z-scores at the 3′ and 5′ ends of the genome (data not shown).

DISCUSSION

The informationally complex HCV genome can essentially be regarded as functioning at three levels. The single-stranded, plus-sense RNA molecule is translated in the cytoplasm to yield viral proteins; it is used as a template for negative-strand synthesis; and it interacts with viral structural components to yield progeny viral particles. These three processes may all rely on secondary (and in some cases tertiary) RNA structures that reside within the HCV genome. Although such structure has been demonstrated within the HCV 5′ and 3′ untranslated regions, little progress has been made toward the identification and characterization of structures residing within the HCV coding sequence. In this study, we identified such regions of genotypically conserved secondary structure in the viral genomic RNA using a combination of established thermodynamic prediction models and newly developed phylogenetic methods that exploit the vast amount of comparative sequence data for HCV.

Detection of RNA secondary structure

A variety of physical and computational methods have been developed to determine RNA secondary structure in viruses and other organisms. In this investigation, the length of the HCV genomic sequence (9,400 bases) prevented the use of physical methods such as nuclear magnetic resonance spectroscopy, enzymatic cleavage, or chemical degradation. Although it would have been possible to separately analyze subgenomic fragments of HCV sequence by RNAse mapping or other probing techniques, the likely complexity of the RNA secondary structure and the possible dependence on long-range interactions for folding would largely invalidate attempts to build up an overall structure from the sum of those determined from short and arbitrarily truncated HCV RNA transcripts. For example, short fragments of RNA suitable for RNAse mapping (100–200 bases) that contained nt 47 to 84 or 8376 to 8455 would inevitably fold to form the stem-loops SL47 and SL8376, but that would not constitute evidence that they existed in full-length genomic RNA.

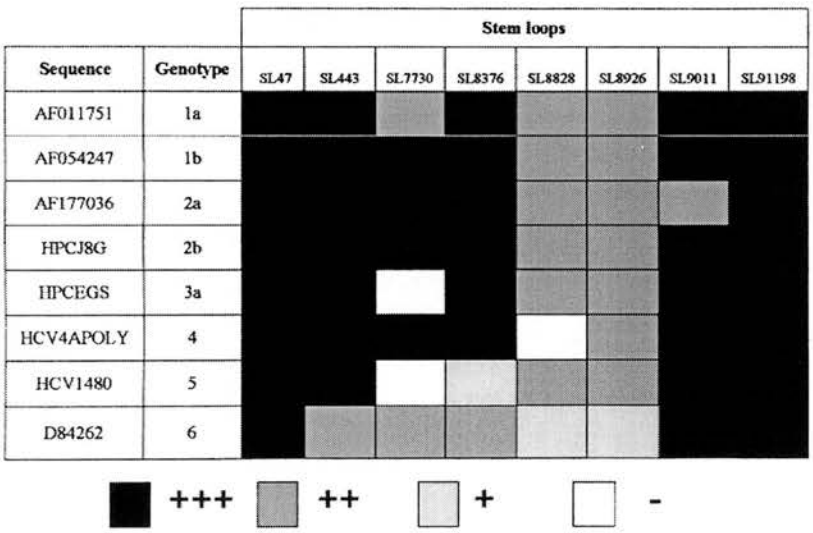


FIGURE 10. Structure conservation of predicted stem-loops in the core and NS5B regions between representative sequences of HCV genotypes 1–6. Structures were scored from − to +++ depending on the degree of similarity to the most common structure as follows: +++: Stem-loop structurally identical; ++: minor differences in base pairing but conservation of overall size and shape of stem-loop; +: different structure in the same region; −: no secondary structure detected.

In contrast, many of the methods used in computational analysis of RNA structure, such as covariance detection and measurement of synonymous variability, were not impeded by sequence length. Indeed, the analysis of HCV was aided considerably by the availability of the large amount of comparative sequence information of different HCV genotypes and variants within genotypes or subtypes. This allowed us to incorporate phylogenetic information into the predictions of base pairing, and, in particular, to apply a covariance scanning algorithm to alignments of HCV genome sequences. For sequences with a marked phylogenetic structure such as HCV, the program represents a substantial improvement over that used previously to analyze HGV/GBV-C sequences (Simmonds & Smith, 1999) because it is better able to reconstruct the evolutionary history of highly variable sites through the use of parsimony.

The data set of published HCV sequences also allowed analyses of variability at synonymous sites in the HCV coding region. Synonymous variability cannot alter the virus phenotype through changes in the encoded proteins, and it had generally been considered that variability is selectively neutral, subject to relatively minor constraints arising from biased codon choice and base composition differences. The observed differences in synonymous sequence heterogeneity in the regions of the respective ORFs of HCV, HGV/GBV-C, and picornaviruses are therefore likely to indicate selection pressures unrelated to their coding function. In the case of HGV/GBV-C, we previously found that suppression of synonymous variability acted as a signature for extensive RNA secondary structure formation in the coding part of the genome that was predicted by independent computational methods (Simmonds & Smith, 1999). As another example of this association, the marked dip in synonymous variability observed on comparing enterovirus coding sequences (Fig. 2D) localized precisely to the *cis*-acting element necessary for strand initiation in poliovirus replication (Goodfellow et al., 2000; Rieder et al., 2000).

We supplemented the standard thermodynamic predictions of RNA secondary structure by comparison of free energies on folding with sequence-order-randomized controls (Fig. 5). Previously published analyses of eukaryotic and prokaryotic gene sequences has revealed a number of potential causes of artefactual results from such methods. These include the effect of regional differences in base composition (such as GC islands) on the calculation of folding free energy differences between native and sequence-order-randomized controls (Rivas & Eddy, 2000). This was ruled out as a cause for the observed free energy differences observed in HCV sequences for two reasons. First, base composition was relatively homogeneous throughout the HCV genome, with a relatively constant moderate overrepresentation of G/C residues at third base positions (Fig. 3A). Second, we developed codon-swapping methods for sequence randomization that minimized the distance between sites that were shuffled, and therefore prevented any local base composition differences from being disturbed (CLS). Such methods produced similar folding free energy differences from methods which fully randomized codon order (COR, CLR; Fig. 5).

A second cause for artefactual folding free energy differences arises from the disruption of dinucleotide pairs by standard sequence scrambling methods (Workman & Krogh, 1999). In this study, we developed two sequence randomization methods that retained the dinucleotide frequencies as well as codon structure in the native sequence, and additionally in the case of CDS, any regional differences there might be in dinucleotide composition. As described above, these latter methods produced free energy differences remarkably similar to those produced by methods that disrupted dinucleotides (Fig. 5). Further evidence that disruption of dinucleotide frequencies was not responsible for the folding free energy differences in HCV (and HGV/GBV-C) sequences was provided by the observation for relatively limited biases in dinucleotide frequencies in HCV sequences; only the CG and UG dinucleotide frequencies were different from those expected from the local base composition. Furthermore, these biases were distributed throughout the HCV genome (Fig. 3B), and did not localize specifically to the 5′ and 3′ ends where the largest folding free energy differences were observed (Fig. 2A, B).

Folding free energy differences observed for HCV were comparable to those observed for viral sequences with well-defined RNA secondary structure, such as the noncoding region of hepatitis delta virus and plant viroids, where free energy differences of between 15–25% have been previously reported using the NOR method for sequence scrambling (Cuceanu et al., 2001). Further evidence for the veracity of the HCV results is provided by the consistent absence of folding free energy differences of four different mammalian coding sequences with any of the scrambling methods (Fig. 5). Indeed, the observation that these sequences show a wide range of base composition differences and more extreme biases in dinucleotide frequencies provided further evidence that these factors have no significant influence on the folding free energies determined in the current study.

The observation of folding free energy differences throughout the HCV genome therefore leads to the remarkable conclusion that RNA secondary structure may be distributed throughout the viral coding sequence. Indeed, the observation that folding free energy differences of reverse complemented sequences were consistently lower than those from the plus strand of RNA indicates a greater likelihood that the RNA structure is more relevant functionally for viral RNA sequences rather than (antisense) replication intermediates. Rather than invalidating our own study, the previous observations

that folding free energy differences are generally incapable of detecting structured RNA elements in eukaryotic and prokaryotic genome sequences (Workman & Krogh, 1999; Rivas & Eddy, 2000) likely provides a preliminary indication of the great organizational difference in RNA sequences of HCV (and HGV/GBV-C) viral RNA from mRNAs and other RNA elements in their host cells.

Location of RNA structure in HCV

This study is the first comprehensive evaluation of secondary structure in the coding region of HCV using a number of independent computational methods. Covariance scanning, thermodynamic predictions, and synonymous variability concurred in the prediction of conserved folded RNA structure elements in the core- and NS5B-encoding regions of the genome. The results confirm the existence of a number of structures predicted independently either by simple sequence inspection (Han & Houghton, 1992; Smith & Simmonds, 1997), or by a comparative RNA-folding algorithm that identified covariant sites through the comparison of phylogenetically conserved RNA structures (Hofacker et al., 1998). The latter study predicted the existence of SL7729 and the terminal region of SL8828.

Predictions of stem-loops by covariance scanning and suppression of synonymous variability detected secondary structures conserved between HCV genotypes (Figs. 6, 7, and 8). However, comparison of individual structures indicated some differences in the extent of folding, and, in some cases, in the identity of the actual bases involved in pairing interactions in the predicted stems (Fig. 10). Similar structural differences have previously between observed between human and chimpanzee HGV/GBV-C sequences (Cuceanu et al., 2001), and suggest some flexibility in whatever functional requirement there may be for such structures (see below). Although it is possible that many other, nonconserved structures exist elsewhere in the HCV coding region, the trend for the greatest differences in free energy to be found at the ends of the genome (Fig. 7) suggest that most folding is concentrated in regions where conserved structures are found.

Comparison with other RNA viruses

The distribution of synonymous variability between genotypes or serotypes of other positive-stranded RNA viruses showed remarkable contrasts to HCV. Using a similar range of thermodynamic and phylogenetic prediction methods, we previously found that HGV/GBV-C showed extensive, conserved RNA structural elements (Simmonds & Smith, 1999; Cuceanu et al., 2001). Although it is possible that the much greater sequence diversity of HCV genotypes prevented the detection of nonconserved stem-loop structures by phylogenetic methods, HCV sequences also showed less difference in free energy on folding than sequence-order-randomized controls (see above). Why HCV secondary structure formation should be less extensive than in HGV/GBV-C is currently unclear.

An even greater contrast is found on analysis of pestivirus sequences. Despite their similarity in genome organization to HCV, there was no evidence for suppression of synonymous variability in any region of the latter's coding sequences. A lack of secondary structure in pestiviruses was also indicated by the much lower difference in free energy on folding BVDV or CSFV sequences with sequence-order-randomized controls (mean values over coding part of the genome: 3.2% and 2.1%; A. Tuplin & P. Simmonds, unpubl. data). Finally, covariance scanning failed to predict any covariant sites between variants of BVDV, BDV, or CSFV sequences (data not shown). The lack of evidence of secondary structure is possibly reflected in the recent finding that IRES-driven translation of BVDV coding sequences is strongly inhibited by base pairing downstream of the methionine initiating codon, for example, by placing the IRES upstream from the GC-rich NS3 sequence of BVDV (Myers et al., 2001). This finding is different from the requirement for structured RNA sequences in the core region of HCV for efficient translation (Reynolds et al., 1995; Honda et al., 1996b; Lu & Wimmer, 1996).

It could be argued that some degree of secondary structure is required to generate the compactness of the RNA genome to assist viral packaging or other replicative steps. The existence of sequence-order-dependent structures in HCV and HGV/GBV-C may therefore compensate for a particular base composition that prevents tight packing of RNA. However, if GC content is a guide to the likelihood on internal base pairing, then the converse appears to be true; The GC content of pestiviruses is only 46%–47%, compared with 58% and 59% for HCV and HGV/GBV-C, respectively.

Role of RNA secondary structure

A total of eight RNA structures were identified in this study. These predictions add to the list of structural elements residing in the coding sequences of other positive-sense RNA viruses. These include the *cis*-acting replicating elements in poliovirus (Goodfellow et al., 2000), in the cardioviruses Theiler's murine encephalomyelitis virus and Mengo virus (Lobert et al., 1999), in human rhinovirus type 14 (HRV-14; McKnight & Lemon, 1998), and in mouse hepatitis virus strain JHM (Kim & Makino, 1995). These *cis*-acting elements have been characterized and shown to be involved in viral replication (Kim & Makino, 1995; McKnight & Lemon, 1998; Lobert et al., 1999; Goodfellow et al., 2000). Although the function of the RNA structures we have identified within the HCV genome remains to be

elucidated, their possible involvement/requirement for translation and/or replication can now be investigated using the recently developed HCV subgenomic RNA-replicon system (Lohmann et al., 1999; Blight et al., 2000). However, a role of the predicted structures in encapsidation of HCV RNA during virion assembly would require the development of a packaging assay or a replicating HCV clone that produced virus particles.

The RNA structures identified prior to the stop codon in the NS5B sequence (SL9011 and SL9118) possibly represent a 5′ extension to the wealth of structure contained within the 3′ UTR. Because this 5′ boundary to the 3′ terminal RNA structures of the HCV genome encroaches into the C-terminus of the polyprotein, the concept that the structure and function of the 3′ terminus of the virus genome is solely contained within the 3′ UTR could perhaps now be considered outmoded. The identification of such structure within this region could also facilitate the elucidation of the 3′ terminus function, which is currently poorly understood. Besides a role in translational regulation (Ito et al., 1998; Ito & Lai, 1999), the 3′ UTR is absolutely required for infectivity in the chimpanzee animal model (Yanagi et al., 1999), and has recently been shown to bind the helicase domain of NS3 (Banerjee & Dasgupta, 2001), and is thus presumed to be involved in viral replication. Of the two RNA structures identified in the core coding sequence (SL47 and SL443), it is interesting to note that SL443 lies 5′ of a pyrimidine-rich domain and putative PTB-binding site previously implicated in HCV translational regulation (Ito & Lai, 1999). Additionally, SL443 may also be involved in translational regulation through a long-range RNA–RNA interaction with sequences at the 5′ end of the genome (Honda et al., 1999).

Apart from secondary structure, suppression of synonymous variability at synonymous sites may occur in sequences translated in alternative reading frames; several investigators have proposed that core gene of HCV may encode a second protein in the +1 reading frame (Ina et al., 1994; Walewski et al., 2001; Xu et al., 2001). Proteins translated from HCV RNA in vitro included a polypeptide of approximately 160 amino acids, containing the first 11 codons of the core gene, and the remaining amino acids encoded by the +1 reading frame (Xu et al., 2001). The authors identified a "slippery" homopolymeric tract of A residues upstream from the proposed −2 ribosomal frameshift site. Frameshifting generally requires the presence of downstream RNA secondary structure, often in the form of a pseudoknot, to pause translation and facilitate transfer of reading frame (Brierley et al., 1992; Matsufuji et al., 1996; Giedroc et al., 2000). Although the authors do not identify an RNA structure in the HCV core gene, an obvious candidate would be SL47, which lies in an appropriate position in relation to the homopolymeric tract (5 nt downstream) to promote frameshifting (Xu et al., 2001). However, assigning a functional role to the product of the +1 reading frame (the F protein) is made problematic by the lack of evolutionary conservation of the coding sequence (Smith & Simmonds, 1997). Approximately half of the published sequences of different HCV genotypes contain premature stop codons at a range of different nucleotide positions in the core sequence (e.g., at 377, 419, 431, and 464), and would therefore produce a heterogeneous range of F proteins, truncated at various positions at the carboxyl terminus of the protein. Much shorter potential coding sequences are found in the sequences HPCHK6 (genotype 3a) and VN405 (genotype 8b); stop codons at positions 86 and 37 would encode polypeptides with predicted lengths of only 29 and 22 amino acids, respectively. As the proposed F protein is clearly dispensable in HPCHK6 and VN405, there is little likelihood that selection pressure to maintain the reading frame in other HCV variants underlies the observed suppression in synonymous variability in the part of the HCV genome.

Beyond RNA–RNA interactions, association of any of the eight RNA structures with viral or cellular proteins is also likely. Identification of such factors is possible through techniques such as yeast three-hybrid screening (SenGupta et al., 1996), as recently reported for the HCV 3′X (Wood et al., 2001), or using proteomic analysis in conjunction with cells stably expressing HCV replicons. The exact structures of the RNA stem-loops identified in this study could be equivocally established using chemical and enzymatic cleavage analysis, although the improvement in RNA secondary structure prediction algorithms combined with the fact that these stem-loops are small and discrete probably mean that there would be little discrepancy between actual and predicted structure.

In summary, we have identified eight genotypically conserved RNA structures that reside within the HCV coding sequence. The role of RNA secondary and tertiary structure in governing essential viral processes is becoming increasingly obvious. The identification of these RNA structures in conjunction with structures known to exist within the HCV untranslated regions may facilitate further understanding of HCV translation, replication, and packaging, or at least may provide an insight into a previously unapparent level of functional and evolutionary complexity residing within the HCV RNA genome.

MATERIALS AND METHODS

Genome sequences

Epidemiologically unlinked complete genome sequences of HCV analyzed in the study were as follows (GenBank accession number in parentheses if different from entry name): HPCPLYPRE (M62321), H77 (AF011751), HC-J1 (D10749), HEC278830 (AJ278830), HC-J4 (AF054247), BK (AF33324),

HPCGENANTI (M84754), HPCCGENOM (L02836), HCU01214 (U01214), HCV-J (NC_001433), HD-1 (U45476), HPCRNA (D10934), HCVJK1G (X61596), JTB (D11355), HPVHCVN (D63857), PP (D30613), HCV-N (S62220), HPCUNKCDS (M96362), HC-G9 (D14853), HC-J6 (AF177036), NDM228 (AF169002), G2aK1 (AF169003), G2aK3 (AF169004), NDM59 (AF169005), HC-J8 (D10988, D01221), BEBE1 (D50409), HPCK3A (D28917), HPCEGS (D17763), HPCFG (D49374), JK049 (D63821), HCV4APOLY (Y11604), AF064490, HCV1480 (Y13184), HCV12083 (Y12083), Th580 (D84262), VN235 (D84263), VN405 (D84264), VN004 (D84265), and JK046 (D63822). Underlined sequences (of genotypes 1a, 2a, 3a, 4a, 5a, and 6a) were use for free energy predictions. For the extended analysis of type 1b sequences, the following additional sequences were compared: AB049087, AB049088, AB049089, AB049091, AB049092, AB049093, AB049094, AB049095, AB049096, AB049098, AB049099, AB049100, AB049101, AF165046, AF165048, AF165050, AF165052, AF165054, AF165056, AF165058, AF165060, AF165062, AF165064, AF176573, AF207752, AF207753, AF207754, AF207756, AF207758, AF207760, AF207761, AF207762, AF207763, AF207764, AF207765, AF207766, AF207767, AF207768, AF207771, AF207772, AF208024, D85516, D89815, D89872, HCJ238800 (AJ238800), HCV132997 (AJ132997), HCVPOLYP (AJ000009), HPC1B4 (D50484), HPC1B5 (D50485), HPCJRNA (D14484, D001173), and HPCY1B6 (D50480).

For comparison, the following sequences of GBV-B, pestiviruses, and enteroviruses were analyzed: GBV-B: NC001655; pestiviruses: BVDCG (M31182), BVDPOLYPRO (M96751), AF091605, BVDPP (M96687), PTU86600 (U86600), NC_002514, AF037405, BDU70263 (U70263), AF002227, HCVPOLYPR (Z46258), AF091507, HCVPOLYP2 (D49533), AF091661, HCVPOLYP1 (D49532), HCU45478 (U45478), HCVSEQB (L49347), A16790, HCVCG3PE (M31768), NC_002657, AF099102, and AF092448; enteroviruses: NC_002058, NC_002029, POL2LAN (M12197), NC_001428, NC_001429, NC_001342, NC_002485, AF231765, NC_000881, NC_000873, NC_002003, NC_001657, NC_001656, NC_001360, NC_001472, NC_002601, NC_001612, AF304459, NC_002347, NC_000945, and NC_001430.

Coding regions from the following mammalian sequences were used as negative controls: actin (BC015695); albumin (AF116645); HLA DRw12 beta 1-chain; and alphaglobin (V00493).

Analysis of synonymous sequence variability

Synonymous sequence variability was determined by parsimony for each codon in alignments of HCV, HGV/GBV-C, pestivirus, and enterovirus complete genome sequences. The phylogeny and sequences of ancestral nodes for each sequence alignment were determined by the program DNAPARS in the PHYLIP package (Felsenstein, 1993). Variability at each codon was expressed as the proportion of comparisons between each sequence or node with the reconstructed codon of its immediate ancestor showing synonymous differences. This method was chosen over simple pairwise comparison at each codon position (Simmonds & Smith, 1999), as information on phylogeny is more effective at reconstructing the likely multiple substitution events found in the highly divergent HCV, pestivirus, and enterovirus sequences. Variability at each codon was normalized to allow differences in variability at saturation at codon comparisons with two-, three-, four-, and sixfold degeneracy. Variability was calculated only at codon positions where 40% or greater of sequence/ancestor comparisons were synonymous. Variability at each codon position was averaged over a sliding window of 50 codons.

Detection of covariance

An alignment of the coding regions of 41 complete genome sequences of HCV was analyzed for covariant changes. The method used for scoring covariance was modified from that previously used to analyze HGV/GBV-C sequences (Simmonds & Smith, 1999) to allow for the tree structure of HCV, and in particular, the nonindependence of covariant changes found in members of individual clades. Before scanning, the phylogeny and ancestral sequences were determined by DNAPARS. To score covariance, each sequence or node was compared with its immediate ancestor to determine the number of evolutionary events at each paired site. This approach avoids the problem of multiply scoring the same covariant substitution found in members of the descendant clade. This was identified as a particular problem with the analysis of HCV sequences, which show several tiers of sequence variability (genotype, subtype, isolate).

Free energy calculations

Coding regions of aligned HCV sequences of genotypes 1a (HPCPLYPRE), 1b (HC-J1), 2a (HC-J6), 3a (HPCEGS), 4a (HCV4APO), 5a (HCV1480), 6a (HCV12083), and of GBV-B (NC_001655) were split into 500-base fragments overlapping by 250 bases. The free energy of folding was calculated using the program MFOLD (Mathews et al., 1999) using default settings. The contribution of nucleotide order to free energy of folding was estimated by comparison of free energy with the mean value of sequences generated by sequence order randomizations. Six different methods were used to randomize coding sequence order as described in Results [nucleotide order randomization (NOR), codon order randomization (COR), like-codon randomization (CLR), like-codon swap (CLS), dinucleotide randomization (CDR), and dinucleotide swap (CDS)]. Free energy results were expressed as the ratio of free energy on folding native sequences to that of the sequence randomized by one of the six methods. For methods NOR, COR, CLR, and CDR, differences in free energy between native and randomized sequences were also be expressed as a *Z*-score, as previously described (Workman & Krogh, 1999). *Z*-scores are the number of standard deviations by which the predicted free energy of the native sequence is lower than the mean of the randomized sequences.

Specific predictions of RNA secondary structure were made for regions of the HCV genome showing suppression of synonymous substitutions, covariant sites associated with stem-loop structure, and excess free energy on folding compared with sequence-order-randomized controls. In practice, this

included the core- and NS5B-encoding regions of HCV. Conservation of each predicted structure was assessed by parallel folding of sequences of different HCV genotypes, and by retention of specific structure within the different structural predictions using different free energy parameters. RNA stem-loop structures have been referred to provisionally in this study by the base position in the HCV alignment of the first base at the 5′ end of the base-paired region. The labeling does not constitute a specific proposal for their future nomenclature.

Sequence software

All free energy calculations and secondary structure predictions were made using the program MFOLD with default settings. Sequence alignments, measurement of synonymous variability, sequence order randomization, measurement of base composition and dinucleotide frequencies, and covariance scanning by parsimony were performed with the Simmonic 2000 package (Simmonds & Smith, 1999), which is available from the authors.

ACKNOWLEDGMENTS

We thanks Dr. Michael Zuker for providing extensive use of the MFOLD server used to calculate folding free energies of the >100,000 sequences analyzed in this study.

Received January 22, 2002; returned for revision April 2, 2002; revised manuscript received April 5, 2002

REFERENCES

Banerjee R, Dasgupta A. 2001. Specific interaction of hepatitis C virus protease/helicase NS3 with the 3′-terminal sequences of viral positive- and negative-strand RNA. *J Virol 75*:1708–1721.

Blight KJ, Kolykhalov AA, Rice CM. 2000. Efficient initiation of HCV RNA replication in cell culture. *Science 290*:1972–1974.

Brierley I, Jenner AJ, Inglis SC. 1992. Mutational analysis of the "slippery-sequence" component of a coronavirus ribosomal frameshifting signal. *J Mol Biol 227*:463–479.

Choo QL, Kuo G, Weiner AJ, Overby LR, Bradley DW, Houghton M. 1989. Isolation of a cDNA derived from a blood-borne non-A, non-B hepatitis genome. *Science 244*:359–362.

Cuceanu NM, Tuplin A, Simmonds P. 2001. Evolutionarily conserved RNA secondary structures in coding and non-coding sequences at the 3′ end of the hepatitis G virus/GB-virus C genome. *J Gen Virol 82*:713–722.

Felsenstein J. 1993. *PHYLIP Inference Package*, version 3.5. Department of Genetics, University of Washington, Seattle.

Giedroc DP, Theimer CA, Nixon PL. 2000. Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting. *J Mol Biol 298*:167–185.

Goodfellow I, Chaudhry Y, Richardson A, Meredith J, Almond JW, Barclay W, Evans DJ. 2000. Identification of a *cis*-acting replication element within the poliovirus coding region. *J Virol 74*:4590–4600.

Han JH, Houghton M. 1992. Group specific sequences and conserved secondary structures at the 3′ end of HCV genome and its implication for viral replication. *Nucleic Acids Res 20*:3520.

Hofacker IL, Fekete M, Flamm C, Huynen MA, Rauscher S, Stolorz PE, Stadler PF. 1998. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucleic Acids Res 26*:3825–3836.

Honda M, Brown EA, Lemon SM. 1996a. Stability of a stem-loop involving the initiator AUG controls the efficiency of internal initiation of translation on hepatitis C virus RNA. *RNA 2*:955–968.

Honda M, Ping LH, Rijnbrand RCA, Amphlett E, Clarke B, Rowlands D, Lemon SM. 1996b. Structural requirements for initiation of translation by internal ribosome entry within genome-length hepatitis C virus RNA. *Virol 222*:31–42.

Honda M, Rijnbrand R, Abell G, Kim DS, Lemon SM. 1999. Natural variation in translational activities of the 5′ nontranslated RNAs of hepatitis C virus genotypes 1a and 1b: Evidence for a long-range RNA–RNA interaction outside of the internal ribosomal entry site. *J Virol 73*:4941–4951.

Ina Y, Mizokami M, Ohba K, Gojobori T. 1994. Reduction of synonymous substitutions in the core protein gene of hepatitis C virus. *J Mol Evol 38*:50–56.

Ito T, Lai MMC. 1999. An internal polypyrimidine-tract-binding protein-binding site in the hepatitis C virus RNA attenuates translation, which is relieved by the 3′-untranslated sequence. *Virol 254*:288–296.

Ito T, Tahara SM, Lai MMC. 1998. The 3′-untranslated region of hepatitis C virus RNA enhances translation from an internal ribosomal entry site. *J Virol 72*:8789–8796.

Kim YN, Makino S. 1995. Characterization of a murine coronavirus defective interfering RNA internal *cis*-acting replication signal. *J Virol 69*:4963–4971.

Kuo G, Choo QL, Alter HJ, Gitnick GL, Redeker AG, Purcell RH, Miyamura T, Dienstag JL, Alter MJ, Stevens CE, Tegtmeier F, Bonino F, Columbo M, Lee W-S, Kuo C, Berger K, Schuster JR, Overby LR, Bradley DW, Houghton M. 1989. An assay for circulating antibodies to a major etiologic virus of human non-A, non-B hepatitis. *Science 244*:362–364.

Lobert PE, Escriou N, Ruelle J, Michiels T. 1999. A coding RNA sequence acts as a replication signal in cardioviruses. *Proc Natl Acad Sci USA 96*:11560–11565.

Lohmann V, Korner F, Koch JO, Herian U, Theilmann L, Bartenschlager R. 1999. Replication of subgenomic hepatitis C virus RNAs in a hepatoma cell line. *Science 285*:110–113.

Lu HH, Wimmer E. 1996. Poliovirus chimeras replicating under the translational control of genetic elements of hepatitis C virus reveal unusual properties of the internal ribosomal entry site of hepatitis C virus. *Proc Natl Acad Sci USA 93*:1412–1417.

Mathews DH, Sabina J, Zuker M, Turner DH. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol 288*:911–940.

Matsufuji S, Matsufuji T, Wills NM, Gesteland RF, Atkins JF. 1996. Reading two bases twice: Mammalian antizyme frameshifting in yeast. *EMBO J 15*:1360–1370.

McKnight KL, Lemon SM. 1998. The rhinovirus type 14 genome contains an internally located RNA structure that is required for viral replication. *RNA 4*:1569–1584.

Myers TM, Kolupaeva VG, Mendez E, Baginski SG, Frolov I, Hellen CU, Rice CM. 2001. Efficient translation initiation is required for replication of bovine viral diarrhea virus subgenomic replicons. *J Virol 75*:4226–4238.

Reynolds JE, Kaminski A, Carroll AR, Clarke BE, Rowlands DJ, Jackson RJ. 1996. Internal initiation of translation of hepatitis C virus RNA: The ribosome entry site is at the authentic initiation codon. *RNA 2*:867–878.

Reynolds JE, Kaminski A, Kettinen HJ, Grace K, Clarke BE, Carroll AR, Rowlands DJ, Jackson RJ. 1995. Unique features of internal initiation of hepatitis C virus RNA translation. *EMBO J 14*:6010–6020.

Rieder E, Paul AV, Kim DW, van Boom JH, Wimmer E. 2000. Genetic and biochemical studies of poliovirus *cis*-acting replication element cre in relation to VPg uridylylation. *J Virol 74*:10371–10380.

Rivas E, Eddy SR. 2000. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics 16*:583–605.

SenGupta DJ, Zhang B, Kraemer B, Pochart P, Fields S, Wickens M. 1996. A three-hybrid system to detect RNA–protein interactions in vivo. *Proc Natl Acad Sci USA 93*:8496–8501.

Simmonds P, Smith DB. 1999. Structural constraints on RNA virus evolution. *J Virol 73*:5787–5794.

Simons JN, Pilot-Matias TJ, Leary TP, Dawson GJ, Desai SM, Schlauder GG, Muerhoff AS, Erker JC, Buijk SL, Chalmers ML, Vansant CL, Mushahwar IK. 1995. Identification of two flavivirus-like genomes in the GB hepatitis agent. *Proc Natl Acad Sci USA 92*:3401–3405.

Smith DB, Pathirana S, Davidson F, Lawlor E, Power J, Yap PL, Simmonds P. 1997. The origin of hepatitis C virus genotypes. *J Gen Virol 78*:321–328.

Smith DB, Simmonds P. 1997. Characteristics of nucleotide substitution in the hepatitis C virus genome: Constraints on sequence change in coding regions at both ends of the genome. *J Mol Evol 45*:238–246.

Tsukiyama Kohara K, Iizuka N, Kohara M, Nomoto A. 1992. Internal ribosome entry site within hepatitis C virus RNA. *J Virol 66*:1476–1483.

Walewski JL, Keller TR, Stump DD, Branch AD. 2001. Evidence for a new hepatitis C virus antigen encoded in an overlapping reading frame. *RNA 7*:710–721.

Wood J, Frederickson RM, Fields S, Patel AH. 2001. Hepatitis C virus 3′X region interacts with human ribosomal proteins. *J Virol 75*: 1348–1358.

Workman C, Krogh A. 1999. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acid Res 27*:4816–4822.

Xu Z, Choi J, Yen TS, Lu W, Strohecker A, Govindarajan S, Chien D, Selby MJ, Ou J. 2001. Synthesis of a novel hepatitis C virus protein by ribosomal frameshift. *EMBO J 20*:3840–3848.

Yanagi M, St Claire M, Emerson SU, Purcell RH, Bukh J. 1999. In vivo analysis of the 3′ untranslated region of the hepatitis C virus after in vitro mutagenesis of an infectious cDNA clone. *Proc Natl Acad Sci USA 96*:2291–2295.

Journal of General Virology (2001), **82**, 713–722. Printed in Great Britain

Evolutionarily conserved RNA secondary structures in coding and non-coding sequences at the 3′ end of the hepatitis G virus/GB-virus C genome

N. M. Cuceanu,† A. Tuplin and P. Simmonds

Laboratory for Clinical and Molecular Virology, University of Edinburgh, Summerhall, Edinburgh EH9 1QH, UK

Hepatitis G virus (HGV)/GB virus C (GBV-C) causes persistent, non-pathogenic infection in a large proportion of the human population. Epidemiological and genetic evidence indicates a long-term association between HGV/GBV-C and related viruses and a range of primate species, and the co-speciation of these viruses with their hosts during primate evolution. Using a combination of covariance scanning and analysis of variability at synonymous sites, we previously demonstrated that the coding regions of HGV/GBV-C may contain extensive secondary structure of undefined function (Simmonds & Smith, *Journal of Virology* 73, 5787–5794, 1999). In this study we have carried out a detailed comparison of the structure of the 3′untranslated region (3′UTR) of HGV/GBV-C with that of the upstream NS5B coding sequence. By investigation of free energies on folding, secondary structure predictive algorithms and analysis of covariance between HGV/GBV-C genotypes 1–4 and the more distantly related HGV/GBV-C chimpanzee variant, we obtained evidence for extensive RNA secondary structure formation in both regions. In particular, the NS5B region contained long stem–loop structures of up to 38 internally paired nucleotides which were evolutionarily conserved between human and chimpanzee HGV/GBV-C variants. The prediction of similar structures in the same region of hepatitis C virus may allow the functions of these structures to be determined with a more tractable experimental model.

Introduction

A new member of the flavivirus family, variously described as hepatitis G virus (HGV) or GB virus C (GBV-C) has recently been characterized (Linnen *et al.*, 1996; Leary *et al.*, 1996). HGV/GBV-C is widely distributed in human populations, with frequencies of active or past infection ranging from 5 to 15%. Infection is frequently persistent and associated with high levels of circulating viraemia, although there is currently no evidence to link HGV/GBV-C infection to any identifiable hepatic or non-hepatic disease. HGV/GBV-C shows limited genetic heterogeneity, but with marked geographical differences in distribution of the four or five currently classified genotypes. It has been suggested that the presence and subsequent diversification of HGV/GBV-C in humans as they migrated out of Africa 100000 years ago accounts for the current association of different HGV/GBV-C genotypes with particular racial groups (Gonzalez-Perez *et al.*, 1997; Tanaka *et al.*, 1998; Katayama *et al.*, 1997). Supporting this conjecture, sequences from the Far East are almost invariably genotype 3, and this genotype is otherwise only found in native inhabitants of North and South America. In contrast, Caucasian and other populations from India westwards including those of Northern Africa are infected with genotype 2. Genotype 1 is confined to sub-Saharan Africa, and shows the greatest overall sequence diversity (Smith *et al.*, 1997; Muerhoff *et al.*, 1997). As further evidence for a very long-term association of this virus with humans and other primates, viruses closely related to HGV/GBV-C have been found in a variety of Old and New World monkey species, and their phylogenetic relationships mirror those of their hosts. Variants of HGV/GBV-C more divergent than human genotypes can be detected in wild-caught chimpanzees from Central and West Africa (Birkenmeyer *et al.*, 1998; Adams *et al.*, 1998). Even more distantly related viruses, collectively described as GBV-A, have been recovered from

Author for correspondence: Peter Simmonds.
Fax +44 131 650 7965. e-mail Peter.Simmonds@ed.ac.uk

† **Current address:** Centre of Infectious Diseases, Department of Medical Microbiology, Molecular Virology Laboratory, Leiden University Medical Centre, Leiden, The Netherlands.

Table 1. Free energy on folding the NS5B region and 3′UTR of GBV-C/HGV: comparison with sequence order-randomized controls

		NS5B region			3′UTR		
Accession no.	Genotype	ΔG/b*	$\Delta G/b_{sc}$†	% diff.	ΔG/b*	$\Delta G/b_{sc}$†	% diff.
U36380	1	−1·44	−1·19	17·33	−1·58	−1·33	15·91
AB013500	1	−1·51	−1·21	19·87	−1·55	−1·30	16·13
AB013501	2	−1·62	−1·22	24·95	−1·56	−1·29	13·21
U63715	2	−1·54	−1·22	20·78	−1·58	−1·26	20·25
U44402	2	−1·41	−1·20	14·89	−1·50	−1·27	15·33
AB008342	3	−1·41	−1·19	15·60	−1·52	−1·29	15·13
D90601	3	−1·51	−1·20	20·44	−1·51	−1·29	12·81
D87263	3	−1·51	−1·21	19·87	−1·56	−1·33	14·74
AB021287	4	−1·43	−1·17	18·82	−1·48	−1·20	18·92
AB018667	4	−1·43	−1·14	20·28	−1·52	−1·21	20·39
Mean‡		**−1·48 ± 0·07**	**−1·20 ± 0·04**	**18·85 ± 3·37**	**−1·54 ± 0·03**	**−1·26 ± 0·07**	**17·25 ± 3·98**
AF070476	CPZ	−1·47	−1·12	23·80	−1·29	−1·10	14·73
U22303	GBV-A	−1·46	−1·14	21·92	−1·46	−1·26	13·70

* Free energy on folding indicated in kJ/mol per base.
† sc, Sequence order-randomized control sequences (50 for sequences U36380, AB013501, D90601, three for other sequences analysed; mean value shown).
‡ Mean and standard deviation of free energy on folding (kJ/mol per base) or difference from sequence order-randomized controls.

several species of New World monkeys (Bukh & Apgar, 1997; Erker *et al.*, 1998; Leary *et al.*, 1997).

The great genetic stability of HGV/GBV-C and related viruses implied by the evidence for co-evolution with primates is difficult to reconcile with their observed rapid sequence change in individuals over short observation periods (Nakao *et al.*, 1997). We have previously found evidence for constraints on sequence change at many sites in the coding sequence, even at those where substitutions would be synonymous (Simmonds & Smith, 1999). Evidence that RNA secondary structure formation through internal base-pairing limits sequence variability at these sites was provided by the finding of multiple covariant sites spatially associated with potential stem–loop structures amongst HGV/GBV-C sequences of different genotypes. Furthermore, these occurred at positions in the genome that showed reductions in synonymous variability. In that study we excluded non-random nucleotide composition and biased codon usage as compounding factors in the use of RNA folding prediction algorithms and calculation of free energies.

In the current study we have used a variety of phylogenetic and free energy-based predictive algorithms to compare the extent and conservation of RNA secondary structure formation in the 3′ untranslated region (3′UTR) with upstream coding sequences from NS5B, a region encoding the viral RNA polymerase. Our findings indicate that the part of the RNA genome containing the coding sequences is more extensively structured than the 3′UTR, and shows better conservation between variants of HGV/GBV-C infecting different primates. These findings imply an important functional role(s) for the observed secondary structure.

Methods

■ **Sequences.** Currently available complete genomic sequences of HGV/GBV-C genotypes 1–4 (GenBank/EMBL accession numbers in parentheses) which included full-length or near full-length 3′UTR sequences were the genotype 1 sequences GBV-C (U36380) and AB013500; type 2a sequences PNF2161 (U44402), T55875 (AF031827), HGV-1517 (AF31828), HGV-1539 (AF031829), AF121950, HGV-Iw (D87255) and GT110 (D90600); the type 2b sequence GBV-C(EA) (U63715); type 3 sequences GT230 (D90601), HGV-IM71 (AB008342), GSI85 (D87262), D87708–D87714 (Katayama *et al.*, 1998); and type 4 sequences AB018667 and AB021287. Sequences were numbered from the start of the coding region after alignment. The chimpanzee homologue, $HGV/GBV\text{-}C_{CPZ}$, and GBV-A from the New World primate *Sanguinis mystax*, bore the accession numbers AF70476 and U22303. Sequences from the NS5 region of hepatitis C virus (HCV) were obtained from the following published sequences: genotype 1a (M62321), 2a (D00944), 3a (D17763), 4a (Y11604), 5a (Y13184) and 6a (Y12083).

Viroid sequences analysed were citrus exocortis viroid (accession no. X53715), potato spindle tuber viroid (U23058), chrysanthemum chlorotic mottle viroid (AJ247123), Mexican papita viroid (L78463) and potato spindle tuber viroid (X76846). Delta virus sequences of different genotypes were obtained from the following entries: AF098261, AJ000558, M21012, D01075 and M28267. Coding sequences of serum albumin were obtained from the following mammalian species: cat (X84842), cow (Y17729), gerbil (AB006197), horse (X74045), macaque (M90463) and rat (U01222). α-globin coding sequences were obtained from the following mammalian species: baboon (X05289), orang-utan (M12158), duck (X02008), marsupial cat (M17083) and human (V00493).

■ **Additional HGV/GBV-C 3′UTR sequences.** 3′UTR sequences

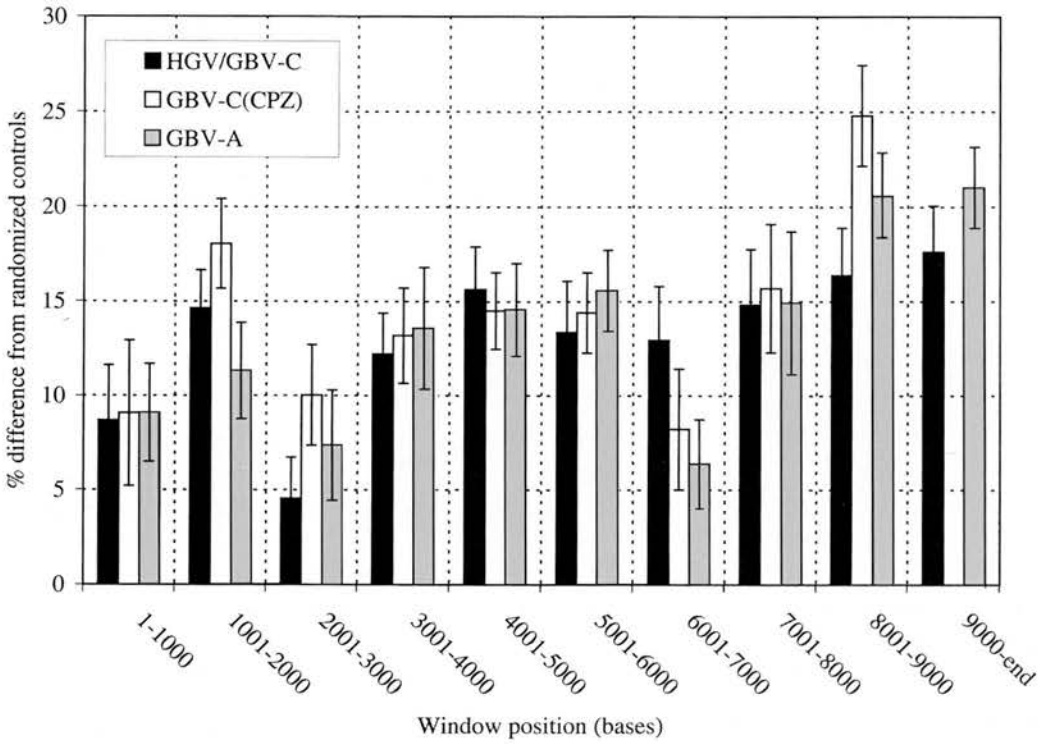


Fig. 1. Free energy on folding consecutive 1000 base fragments of HGV/GBV-C (AB013500), HGV/GBV-C_{CPZ} (AF070476) and GBV-A (U22303) expressed as percentage differences from sequence order-randomized controls. Bars show 95% confidence intervals calculated from multiple sequence randomizations. Bases numbered from start of genome.

Table 2. Free energy on folding of RNA sequences with defined secondary structure and controls

Sequence	*n**	ΔG/b†	$\Delta G/b_{sc}$‡	% diff.
HGV/GBV-C 5′UTR	4	−1·42±0·05	−1·27±0·05	10·27±2·36
Delta virus	5	−1·61±0·04	−1·29±0·01	19·95±1·93
Plant viroid	5	−1·45±0·03	−1·17±0·04	19·47±2·37
α-Globin	5	−1·05±0·14	−1·09±0·12	−2·94±3·22
Albumin	6	−0·80±0·03	−0·82±0·04	−2·30±3·87
HCV NS5B	6	−1·33±0·08	−1·02±0·06	22·55±6·44

* No. of sequences analysed.
† Mean and standard deviation of free energy on folding (kJ/mol per base).
‡ sc, Sequence order-randomized controls (three randomizations per sequence analysed).

were obtained from 17 samples whose genotype had been deduced from sequence comparisons of the 5′UTR and E2 regions (Smith *et al.*, 1997, 2000). RNA was extracted using proteinase K–SDS and phenol–chloroform as described previously (Jarvis *et al.*, 1994). Purified RNA was then reverse-transcribed and amplified by hemi-nested RT–PCR using primers derived from conserved regions of the HGV/GBV-C genome at the carboxyl end of the NS5B gene and the extreme 3′-end: Z3580 – outer sense (positions 8829–8848, 5′ GGTGGTNCATCAATTGGATT 3′, where N = A, C, G, T); Z3581 – inner sense (positions 8881–8900, 5′ GGTTCTTAGCCCTGCTCATC 3′); and Z3582 – outer and inner antisense (positions 9212–9231, 5′ AGTAGAACCCGGCCTTTGGG 3′). Reverse transcription was carried out at 42 °C for 30 min using avian myeloblastosis virus reverse transcriptase (Promega). The conditions for the first round of PCR were hot start at 80 °C for 2 min followed by 30 cycles of 94 °C for 18 s, 58 °C for 21 s and 72 °C for 90 s. At the end of the last cycle, samples were heated to 72 °C for 5 min to allow termination of incomplete strands. The second round of PCR was performed using 1 μl of the primary PCR product for the same number of cycles and conditions. The amplified PCR products were cloned into pGEM-T vector (Promega), and sequenced with both sense and antisense

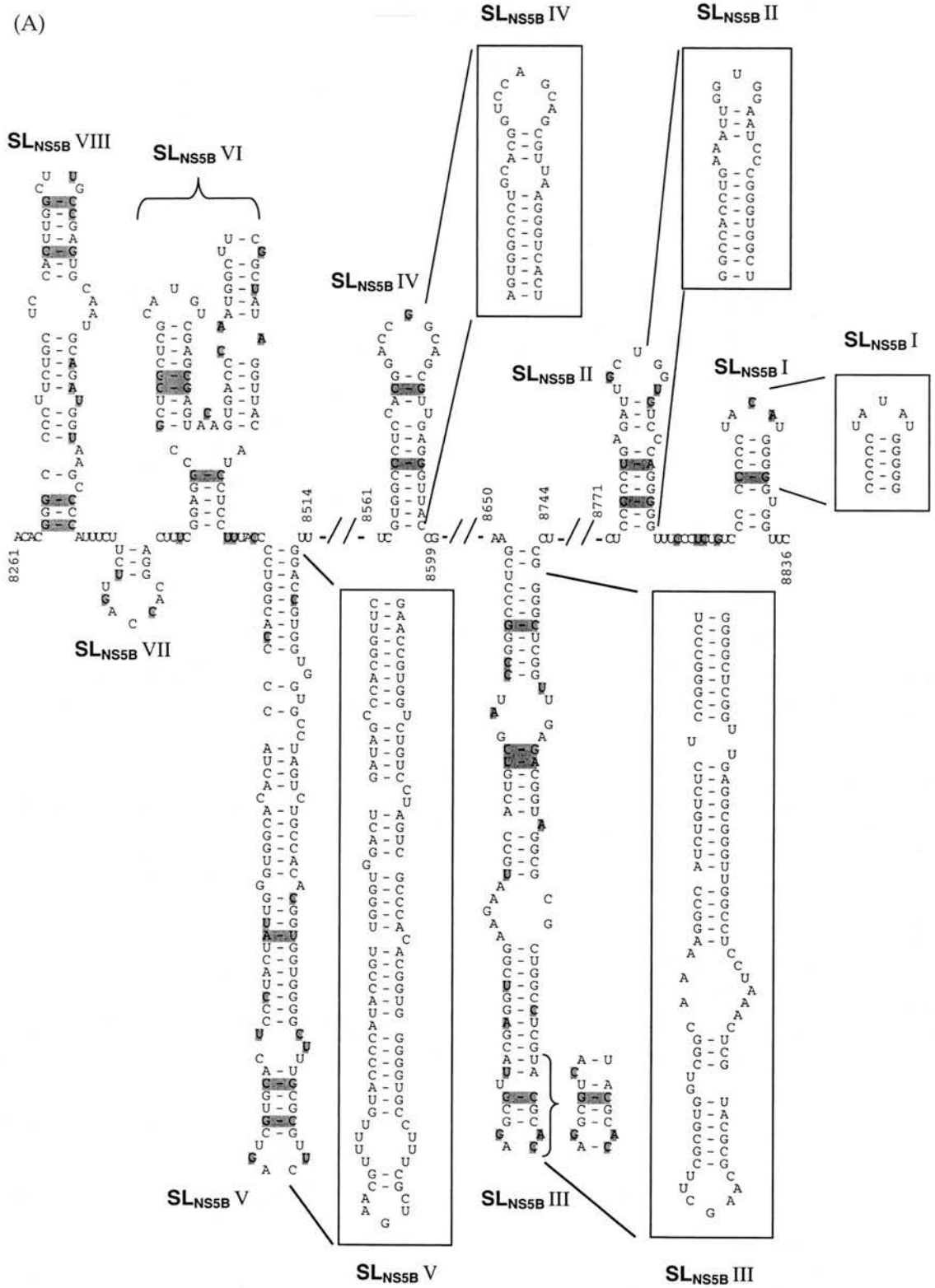


Fig. 2. For legend see facing page.

plasmid primers using T7 DNA polymerase (Sequenase, USB). The consensus sequence of one to three clones for each sample was used for phylogenetic analysis.

■ **Free energy on RNA folding.** The last 1000 bases of aligned HGV/GBV-C complete genome sequences were split into NS5B and 3′UTR regions. The NS5B sequence included the whole sequence

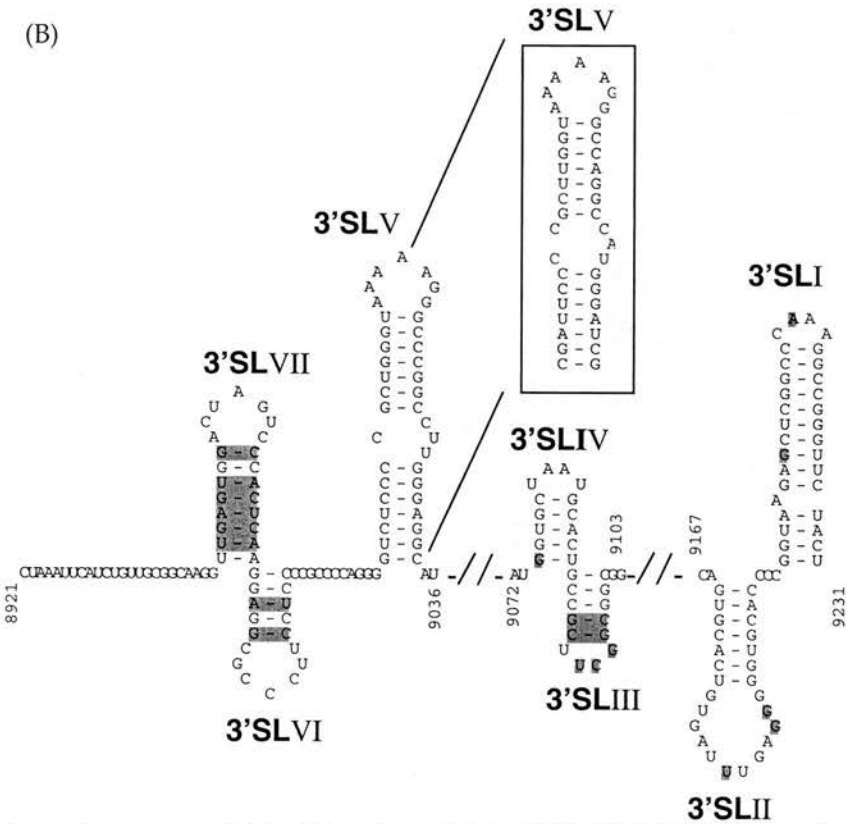


Fig. 2. Predicted secondary structure of (A) the NS5B region and (B) the 3′UTR of HGV/GBV-C. Bases numbered from start of the sequence AB013500. Variable sites between HGV/GBV-C genotypes 1–4 shown in grey boxes. Symbols: '–', canonical Watson–Crick base pairing or GU pairing; '//', intervening sequence without secondary structure (not shown). Stem–loops formed by HGV/GBV-C_{CPZ} that differ in structure shown in boxes.

upstream of and including the stop codon (670 bases in the sequence U36380), while the 3′UTR included the whole sequence downstream from this position (321 bases). The free energy of folding was calculated with the programs RNADraw v1.1 or MFOLD using default settings. The contribution of nucleotide order to free energy of folding was estimated by comparison of free energy with the mean value of sequences generated by independent sequence order randomizations. The variability in free energy on folding 50 sequence order randomizations of three representative HGV/GBV-C sequences of genotype 1 (U36380), 2 (AB013501) and 3 (D90601) was comparable to the combined variability shown by three sequence order randomizations of the seven HGV/GBV-C sequences shown in Table 1 [NS5 region: ±0·044 (±2·7%), ±0·041 (±2·8%) and ±0·040 (±2·8%); 3′UTR: ±0·043 (±4·0%), ±0·041 (±4·3%) and ±0·04 (±3·7%) for the three sequences with 50 randomizations].

This method was also used to analyse whole virus genomes of HGV/GBV-C, HGV/GBV-C_{CPZ} and GBV-A using a sliding window of 1000 bases. The free energies on folding each fragment of HGV/GBV-C, HGV/GBV-C_{CPZ} and GBV-A sequences were compared with those of 15 sequence order randomizations. Variability in the free energies of folding the latter sequences was used to calculate a standard error for the free energy difference estimate.

Free energies in the HGV/GBV-C NS5B and 3′UTR were compared with the mean values of sets of four 5′UTR sequences (U44402, U63715, AB008335 and D87263), five plant viroid sequences, five delta virus sequences, and five albumin and five α-globin sequences from a range of vertebrate species.

■ **Sequence software.** All randomization, free energy calculations and secondary structure predictions were made with the programs RNADraw v1.1 and MFOLD using default settings. Sequence alignments and distance measurements were performed with the Simmonic 2000 package, which is available from the authors.

Results

Free energy on RNA folding

The free energy of folding in different parts of the HGV/GBV-C genome was measured in consecutive 1000 base fragments spanning the genome (Fig. 1). Free energy of folding of the HGV/GBV-C sequence was compared with that of independently generated control sequences whose nucleotide sequence order had been randomized. Although each segment of the HGV/GBV-C genome showed a greater free energy than the control sequences, the difference was most marked at the 3′ end of the genome, with the fragments from 7000–8000, 8000–9000 and 9000–end showing differences ranging from 16 to 18%. To investigate whether the effect of sequence ordering on free energy was conserved amongst viruses related to HGV/GBV C, similar analyses were carried out on sequences from HGV/GBV-C_{CPZ} and GBV-A. Although these viruses display only 75% and ≈ 40% sequence similarity to HGV/GBV-C, they both showed remarkably similar free

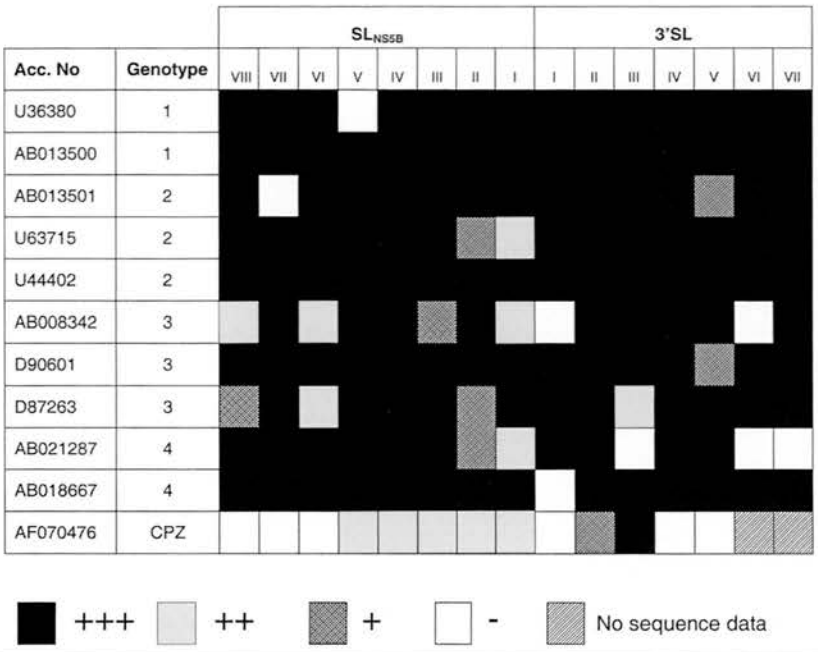


Fig. 3. Structure conservation of predicted stem–loops in the NS5B region (SL_{NS5B}I–SL_{NS5B}VIII) and 3′UTR (3′SLI–3′SLVII) between representative sequences of HGV/GBV-C genotypes depending on the degree of similarity to the most common structure in HGV/GBV-C sequences as follows: '+++', stem–loop structurally identical; '++', minor differences in base-pairing but conservation of overall size and shape of stem–loop; '+', different structure in the same region; '-', no secondary structure detected.

energy profiles, with particularly large differences in the 3′-terminal regions. The large free energies on folding and hence the possibility of secondary structure formation in genomic RNA are therefore a conserved evolutionary feature of this subgenus of viruses.

In this study we have concentrated our analysis on the 3′ terminal region of HGV/GBV-C (Table 1). Fragments of the coding NS5B gene sequences, and the non-coding 3′UTR of different HGV/GBV-C genotypes, showed differences in free energy of 15–23% (mean 18·9 ± 3·4%) and 15–20% (mean 17·3 ± 4·0%) from sequence order-randomized controls. The significance of these free energy differences is indicated by a comparison with sequences of other viruses or virus-like agents with previously documented RNA secondary structure (Table 2). Firstly, the 5′UTR of HGV/GBV-C, whose conserved secondary structure is believed to function as an internal ribosomal entry site (Simons *et al.*, 1996), shows a free energy on folding of 1·4 kJ/base, 10% higher than randomized controls. Secondly, plant viroids contain covalently closed single-stranded RNA genomes with extensive stem–loop structures required for various RNA-catalysed replicative functions. These sequences showed a mean free energy on folding of −1·17 kJ/base (range −1·10 to −1·20), and a difference in free energy of 19·5% (16–23%). Finally, the non-coding region of the single-stranded RNA delta virus genome also contains a well-characterized RNA structure, in which RNA is folded into a ribosome-like domain. The mean free energies on folding [1·29 kJ/base (range −1·27 to −1·32), 20·0% difference from controls (range 15–23%)] of five delta virus variants were also remarkably similar to those calculated for the NS5B region of HGV/GBV-C (Table 2). As negative controls, sequences of the coding regions of several different mammalian and non-mammalian α-globin and albumin genes (which would not be expected to form secondary structures) showed no significant differences in free energy on folding from sequence order-randomized controls. These sequences showed markedly different codon biases; albumin shows a low G + C content (45·1% G + C overall, 34·2% G + C at 3rd base positions), while α-globin has a high G + C content (62·0% overall, 83·0% at 3rd base positions). Since these sequences show no difference from randomized controls in free energy on folding, it is unlikely that biased codon usage per se would influence the values calculated for HGV/GBV-C, as its G + C content lies within these two extremes (63·3% overall, 65·7% at 3rd base positions).

Prediction of RNA secondary structure

Two RNA structure prediction methods (RNADraw and MFOLD) were used to compare the folding of different variants of HGV/GBV-C and the more divergent HGV/GBV-C_{CPZ} (Fig. 2). Results from the two methods were equivalent. Both predicted seven potential stem–loops formed by RNA folding in the 3′UTR (provisionally assigned as 3′SLI–3′SLVII), and eight in the NS5B region (labelled SL_{NS5B}I–SL_{NS5B}VIII). Most of the loops in both regions were conserved between all HGV/GBV-C genotypes, and many were also found in HGV/GBV-C_{CPZ}. A scoring system was adopted to indicate the degree of structure conservation between RNA sequences (Fig. 3), differentiating between identity of RNA structure (+++), similar folding but with minor differences in the positions of bulges and/or minor slippage (++) and folding the same region but with a different structure (+). (Examples of structural differences scored as ++ and + between human and chimpanzee HGV/GBV-C secondary structures are shown

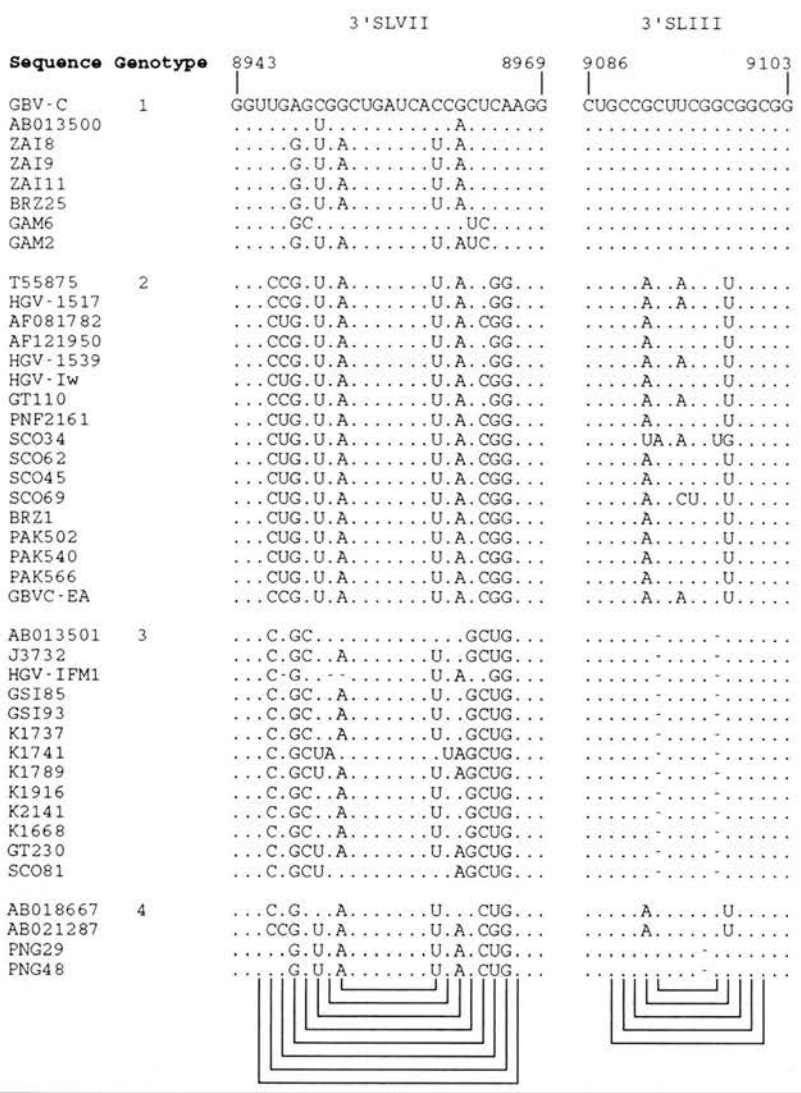


Fig. 4. Covariant substitutions in stem–loops 3′SLVII and 3′SLIII in the 3′UTR between variants of HGV/GBV-C. Bases identical to that of GBV-C indicated as '.'. Proposed base pairings in stem–loops indicated underneath alignment.

in Fig. 2). By these comparisons, substantial structural similarity was found between all sequences in the NS5B region, with similar or identical stem–loops $SL_{NS5B}I$ to $SL_{NS5B}V$ being predicted for both human and chimpanzee HGV/GBV-C sequences. In contrast, only stem–loop 3′SLIII was completely conserved in the 3′UTR, the predicted structure of the HGV/GBV-C_{CPZ} sequence differing considerably elsewhere in this region. The different prediction for this variant may, however, have resulted at least in part from the likely incompleteness of the published HGV/GBV-C_{CPZ} sequence. Alignment of homologous bases indicated that the terminal 55 bases were missing, including the regions forming stem–loops 3′SLII and 3′SLI.

Analysis of covariance

A large number of fully covariant and semi-covariant (C↔U opposite G, A↔G opposite U, and G↔U opposite A) sites were identified in predicted base-paired regions, particularly between human and chimpanzee sequences (Fig. 2). These accommodated the sequence variability between genotypes in structurally conserved regions. Although sequence variability between HGV/GBV-C genotypes was distributed throughout the NS5B and 3′UTR regions, a lower frequency of substitutions were found in regions predicted to be base-paired (18 from 273 sites in NS5B, 2 from 116 sites in the 3′UTR) than in regions predicted to be non-base-paired (57/267 and 21/188; $P < 10^{-7}$ and $P = < 0{\cdot}002$ respectively by Fisher's Exact Test).

One of the problems of verifying structure predictions for the 3′UTR using covariance was lack of comparative sequence information of two of the four HGV/GBV-C genotypes. In GenBank, there are currently only 25 complete or near-complete 3′UTR sequences of which only two are from type 1 and two of type 4. We obtained additional 3′UTR sequences of genotypes 1–4 from HGV/GBV-C-infected individuals from various geographical locations. Sequence variability in the 3′UTR was largely confined to the predicted stem–loop 3′SLVII (Fig. 4), which demonstrated a large number of

covariant substitutions both between and within genotypes. Covariant sites, and in genotype 3 a paired deletion, were also found in stem–loop 3′SLIII. The remaining predicted structures were in regions of the 3′UTR invariant between HGV/GBV-C genotypes, although covariant changes and some structural differences with the HGV/GBV-C_{CPZ} sequence were detected in loops 3′SLVI and 3′SLIV (data not shown).

Sequence variation in the NS5B region was also marked by multiple covariant sites, generally between paired, usually synonymous, 3rd codon positions in predicted stem loops (data not shown). There are also examples of minor structural differences between HGV/GBV-C genotypes, such as the terminal region of stem–loop SL_{NS5B}III. The existence of multiple covariant sites in each of the predicted stem–loops, and their conservation of the majority of structures in the more divergent HGV/GBV-C_{CPZ} sequence, provides strong supporting evidence for the secondary structure predicted by free energy calculations.

Discussion

Structure of the 3′UTR

The secondary structure of the 3′UTR was inferred by folding algorithms based on calculations of free energy, structural conservation between HGV/GBV-C genotypes and the occurrence of covariant changes in base-paired regions. Our prediction differs from that of a previous study which was based on an analysis of the last 140 nucleotides of the 3′UTR from three HGV/GBV-C sequences (Okamoto *et al.*, 1997). The more extensive analysis of 42 sequences in this study identified several substitutions that disrupted the previously proposed base pairings. For example, the proposed pairing of positions 9098 (U) and 9229 (A) would be disrupted in the majority of sequences because of U → C substitution at position 9098, and in one group 2 sequence (SCO34) because of a U → G substitution; another U–A pair between 9109 and 9117 is affected in three sequences (SCO34, PNG29, PNG48) because of a U → C substitution. Other positions at which substitutions produce mismatches in regions predicted to be base-paired by Okamoto *et al.* are: C–C at positions 9140 and 9193 in four group 3 and two unassigned sequences; G–G at positions 9150 and 9189 in three genotype 3 and two genotype 4 sequences; A–G at positions 9130 and 9201 in one genotype 2 sequence. In addition, the Okamoto *et al.* model is not supported by any covariant substitutions and does not include the fragment upstream of position 9092 in which other structures were found.

Computer-predicted folding patterns and RNase cleavage experiments have previously demonstrated the existence of a long stable hairpin structure (3′-LSH) within the distal part of the 3′UTR of several different flaviviruses (Proutski *et al.*, 1997; Brinton *et al.*, 1986; Rice *et al.*, 1985), some positive-strand RNA plant viruses (Strauss & Strauss, 1983), HCV (Blight & Rice, 1997; Kolykhalov *et al.*, 1996), GBV-B (Rijnbrand *et al.*, 2000) and pestiviruses (Yu *et al.*, 1999; Deng & Brock, 1993). Other studies have provided evidence for a specific interaction between the 3′LSH of flaviviruses and host cellular proteins, components of the virus replication complex or have demonstrated a specific binding of cellular proteins to the 3′-terminal 98 nucleotides of the HCV RNA (Ito & Lai, 1997) and determined which regions of the HCV 3′-UTR are critical for *in vivo* virus replication (Lefrere *et al.*, 1999).

The configurations of the predicted terminal loops 3′SLII and 3′SLI in the 3′UTR of HGV/GBV-C (Fig. 2) closely resemble those predicted for HCV (Tanaka *et al.*, 1996; Kolykhalov *et al.*, 1996) and GBV-B (Rijnbrand *et al.*, 2000). However, there was no evidence for a conserved third loop (3′SLIII) nor primary sequence similarity between HGV/GBV-C and HCV or GBV-B. Additionally, the terminal loop is shorter than the HCV and GBV-B homologues (14 base pairs in the stem instead of 19 or 20), although the predicted free energy for formation of the terminal loop (−73 kJ, −2·2 kJ/b) is similar to that of HCV (−110 kJ, −2·4 kJ/b), and indicates a high probability of its formation *in vivo*.

The 3′UTR sequences of human HGV/GBV-C genotypes were highly conserved, with mean pairwise distances between genotypes ranging from 3·8 to 6·6%, compared with 12·8 to 13·4% over the rest of the genome. The HGV/GBV-C_{CPZ} sequence, however, did not display the same differential in divergence, with 23–26% sequence divergence from human genotypes in the 3′UTR, only slightly lower than observed upon comparison of coding sequences (30%). Structurally, only loop 3′SLV was found in both human and chimpanzee variants, although the alignment indicated that the region containing the two terminal loops was missing from the published HGV/GBV-C_{CPZ} sequence. The lack of structural conservation between HGV/GBV-C sequences is not unexpected, given the lack of similarity between other members of the hepaciviruses and pestiviruses in this region. For example, only the terminal three loops are conserved between HCV and GBV-B, and there is also considerable sequence variability between HCV genotypes in non-coding regions 5′ to this, including the great variability in length of the poly(U) tract. Clearly there are varying constraints on sequence change between different regions of the 3′UTR. However, apart from the involvement of the terminal loops in transcription initiation of HCV (Lohmann *et al.*, 1999) and potentially other hepaciviruses, it remains unclear what other functional roles secondary structure in the 3′UTR may play.

Secondary structure in NS5B

The prediction methods used to determine the structure of the 3′UTR were also used to analyse the coding sequence of NS5B. Surprisingly, this region showed even greater free energy on folding than the 3′UTR, several large stem–loops such as those numbered SL_{NS5B}III and SL_{NS5B}V and, in contrast to the 3′UTR, substantial structural similarity between

human and chimpanzee HGV/GBV-C variants. Generally, secondary structures were either identical between HGV/GBV-C variants or, particularly on comparison with the HGV/GBV C_{CPZ} sequence, showed some differences in the identity of the bases involved in base-pairing, but retained conservation of the overall shape and size of the stem–loop. The finding of secondary structure in this region and differences in free energy with sequence order-randomized sequences throughout the genome (Fig. 1) confirms and extends our previous predictions for extensive structure of the HGV/GBV-C genome based on a novel method of covariance scanning and analysis of the distribution of variability at synonymous sites (Simmonds & Smith, 1999). The involvement of such a high proportion of bases in internal base-pairing in the NS5B region, and by implication elsewhere in the genome, suggests that the RNA molecule may be extensively folded through local and possible longer-range interactions to form a 'tertiary' RNA structure.

The conservation in structure and the large number of covariant substitutions suggest a functional role(s) for the RNA structures. Particularly striking was the similarity in free energy on folding the NS5B sequences with the free energies observed for plant viroids and the non-coding region of delta viruses. For these agents, the secondary structure is essential for the replication of the genome, where specific domains may catalyse RNA cleavage, ligation and editing of the genomic RNA sequence. For HGV/GBV-C, amongst several possibilities, RNA folding may be required for packaging of the HGV/GBV-C genome into virus particles, or to protect the genome from RNA-degrading enzymes, particularly as HGV/GBV-C and related viruses do not appear to encode a conventional nucleocapsid protein. In other RNA viruses, secondary structures such as the *cis*-acting replication element in picornaviruses may play a role in initiation of RNA synthesis through long-range interactions with the 3′-terminal region of the genome (Goodfellow *et al.*, 2000; McKnight & Lemon, 1998). The interactions between different genomic regions implied by these observations suggest that HGV/GBV-C might also have an organized overall structure of the RNA genome, in which stem–loop structures may play a role in virus replication.

Enzymatic and chemical methods have been used to provide evidence of secondary structures independently of sequence analysis. While we have considered this approach for the further investigation of the HGV/GBV-C described in this and our previous study (Simmonds & Smith, 1999), the problem with the analysis of sequences such as the NS5B is that they are too long to be easily resolvable by conventional methods. Although it may be possible to separately analyse shorter lengths of sequence in this region, splitting sequences in this way could disrupt the longer-range interactions such as the base-pairing in the base of stem–loops SL_{NS5B}III and V. Direct visualization by electron microscopy of RNA folded in physiological conditions is potentially a better method to determine secondary structure of longer sequences of RNA, particularly if combined with hybridization with gold labelled probes to identify specific sequences within the observed structures. We are currently carrying out this analysis with RNA transcripts from the two regions analysed in this study. Additionally, experimental manipulation of the recently described infectious clone of HGV/GBV-C and methods to culture the virus *in vitro* (Xiang *et al.*, 2000) may allow a direct investigation of the functional significance of RNA folding in this region of the genome.

Using methods described in this and our previous study, we have also commenced secondary structure analyses of other members of the flavivirus family. HCV shows an even greater excess of free energy on folding the NS5B region (23%, Table 2), several stem–loop structures conserved between HCV genotypes (which are much more divergent in nucleotide sequence than between the HGV/GBV-C sequences analysed in this study), and the occurrence of multiple covariant sites in each of the predicted stem–loops (data not shown). The availability of a replicating clone of HCV (Lohmann *et al.*, 1999) may allow the role of such structures to be experimentally investigated.

References

Adams, N. J., Prescott, L. E., Jarvis, L. M., Lewis, J. C. M., McClure, M. O., Smith, D. B. & Simmonds, P. (1998). Detection of a novel flavivirus related to hepatitis G virus/GB virus C in chimpanzees. *Journal of General Virology* **79**, 1871–1877.

Birkenmeyer, L. G., Desai, S. M., Muerhoff, A. S., Leary, T. P., Simons, J. N., Montes, C. C. & Mushahwar, I. K. (1998). Isolation of a GB virus-related genome from a chimpanzee. *Journal of Medical Virology* **56**, 44–51.

Blight, K. J. & Rice, C. M. (1997). Secondary structure determination of the conserved 98-base sequence at the 3′ terminus of hepatitis C virus genome RNA. *Journal of Virology* **71**, 7345–7352.

Brinton, M. A., Fernandez, A. V. & Dispoto, J. H. (1986). The 3′-nucleotides of flavivirus genomic RNA form a conserved secondary structure. *Virology* **153**, 113–121.

Bukh, J. & Apgar, C. L. (1997). Five new or recently discovered (GBV-A) virus species are indigenous to New World monkeys and may constitute a separate genus of the Flaviviridae. *Virology* **229**, 429–436.

Deng, R. T. & Brock, K. V. (1993). 5′ and 3′ untranslated regions of pestivirus genome – primary and secondary structure analyses. *Nucleic Acids Research* **21**, 1949–1957.

Erker, J. C., Desai, S. M., Leary, T. P., Chalmers, M. L., Montes, C. C. & Mushahwar, I. K. (1998). Genomic analysis of two GB virus A variants isolated from captive monkeys. *Journal of General Virology* **79**, 41–45.

Gonzalez-Perez, M. A., Norder, H., Bergstrom, A., Lopez, E., Visona, K. A. & Magnius, L. O. (1997). High prevalence of GB virus C strains genetically related to strains with Asian origin in Nicaraguan hemophiliacs. *Journal of Medical Virology* **52**, 149–155.

Goodfellow, I., Chaudhry, Y., Richardson, A., Meredith, J., Almond, J. W., Barclay, W. & Evans, D. J. (2000). Identification of a cis-acting replication element within the poliovirus coding region. *Journal of Virology* **74**, 4590–4600.

Ito, T. & Lai, M. M. C. (1997). Determination of the secondary structure of and cellular protein binding to the 3′ untranslated region of the hepatitis C virus RNA genome. *Journal of Virology* **71**, 8698–8706.

Jarvis, L. M., Watson, H. G., McOmish, F., Peutherer, J. F., Ludlam, C. A. & Simmonds, P. (1994). Frequent reinfection and reactivation of hepatitis C virus genotypes in multitransfused hemophiliacs. *Journal of Infectious Diseases* **170**, 1018–1022.

Katayama, Y., Apichartpiyakul, C., Handajani, R., Ishido, S. & Hotta, H. (1997). GB virus C hepatitis G virus (GBV-C/HGV) infection in Chiang Mai, Thailand, and identification of variants on the basis of 5′-untranslated region sequences. *Archives of Virology* **142**, 2433–2445.

Katayama, K., Kageyama, T., Fukushi, S., Hoshino, F. B., Kurihara, C., Ishiyama, N., Okamura, H. & Oya, A. (1998). Full-length GBV-C/HGV genomes from nine Japanese isolates: characterization by comparative analyses. *Archives of Virology* **143**, 1063–1075.

Kolykhalov, A. A., Feinstone, S. M. & Rice, C. M. (1996). Identification of a highly conserved sequence element at the 3′ terminus of hepatitis C virus genome RNA. *Journal of Virology* **70**, 3363–3371.

Leary, T. P., Muerhoff, A. S., Simons, J. N., Pilot-Matias, T. J., Erker, J. C., Chalmers, M. L., Schlauder, G. S., Dawson, G. J., Desai, S. M. & Mushahwar, I. K. (1996). Sequence and genomic organization of GBV-C: a novel member of the Flaviviridae associated with human non A–E hepatitis. *Journal of Medical Virology* **48**, 60–67.

Leary, T. P., Desai, S. M., Erker, J. C. & Mushahwar, I. K. (1997). The sequence and genomic organization of a GB virus A variant isolated from captive tamarins. *Journal of General Virology* **78**, 2307–2313.

Lefrere, J. J., Roudotthoraval, F., Morandjoubert, L., Brossard, Y., Parnetmathieu, F., Mariotti, M., Agis, F., Rouet, G., Lerable, J., Lefevre, G., Girot, R. & Loiseau, P. (1999). Prevalence of GB virus type C hepatitis G virus RNA and of anti E2 in individuals at high or low risk for blood borne or sexually transmitted viruses: evidence of sexual and parenteral transmission. *Transfusion* **39**, 83–94.

Linnen, J., Wages, J., Zhangkeck, Z. Y., Fry, K. E., Krawczynski, K. Z., Alter, H., Koonin, E., Gallagher, M., Alter, M., Hadziyannis, S., Karayiannis, P., Fung, K., Nakatsuji, Y., Shih, J. W. K., Young, L., Piatak, M., Hoover, C., Fernandez, J., Chen, S., Zou, J. C., Morris, T., Hyams, K. C., Ismay, S., Lifson, J. D., Hess, G., Foung, S. K. H., Thomas, H., Bradley, D., Margolis, H. & Kim, J. P. (1996). Molecular cloning and disease association of hepatitis G virus: a transfusion-transmissible agent. *Science* **271**, 505–508.

Lohmann, V., Korner, F., Koch, J. O., Herian, U., Theilmann, L. & Bartenschlager, R. (1999). Replication of subgenomic hepatitis C virus RNAs in a hepatoma cell line. *Science* **285**, 110–113.

McKnight, K. L. & Lemon, S. M. (1998). The rhinovirus type 14 genome contains an internally located RNA structure that is required for viral replication. *RNA* **4**, 1569–1584.

Muerhoff, A. S., Smith, D. B., Leary, T. P., Erker, J. C., Desai, S. M. & Mushahwar, I. K. (1997). Identification of GB virus C variants by phylogenetic analysis of 5′ untranslated and coding region sequences. *Journal of Virology* **71**, 6501–6508.

Nakao, H., Okamoto, H., Fukuda, M., Tsuda, F., Mitsui, T., Masuko, K., Lizuka, H., Miyakawa, Y. & Mayumi, M. (1997). Mutation rate of GB virus C hepatitis G virus over the entire genome and in subgenomic regions. *Virology* **233**, 43–50.

Okamoto, H., Nakao, H., Inoue, T., Fukuda, M., Kishimoto, J., Iizuka, H., Tsuda, F., Miyakawa, Y. & Mayumi, M. (1997). The entire nucleotide sequences of two GB virus C/hepatitis G virus isolates of distinct genotypes from Japan. *Journal of General Virology* **78**, 737–745.

Proutski, V., Gould, E. A. & Holmes, E. C. (1997). Secondary structure of the 3′ untranslated region of flaviviruses: similarities and differences. *Nucleic Acids Research* **25**, 1194–1202.

Rice, C. M., Lenches, E. M., Eddy, S. R., Shin, S. J., Sheets, R. L. & Strauss, J. H. (1985). Nucleotide sequence of yellow fever virus: implications for flavivirus gene expression and evolution. *Science* **229**, 726–733.

Rijnbrand, R., Abell, G. & Lemon, S. M. (2000). Mutational analysis of the GB virus B internal ribosome entry site. *Journal of Virology* **74**, 773–783.

Simmonds, P. & Smith, D. B. (1999). Structural constraints on RNA virus evolution. *Journal of Virology* **73**, 5787–5794.

Simons, J. N., Desai, S. M., Schultz, D. E., Lemon, S. M. & Mushahwar, I. K. (1996). Translation initiation in GB viruses A and C: evidence for internal ribosome entry and implications for genome organization. *Journal of Virology* **70**, 6126–6135.

Smith, D. B., Cuceanu, N., Davidson, F., Jarvis, L. M., Mokili, J. L. K., Hamid, S., Ludlam, C. A. & Simmonds, P. (1997). Discrimination of hepatitis G virus/GBV-C geographical variants by analysis of the 5′ non-coding region. *Journal of General Virology* **78**, 1533–1542.

Smith, D. B., Basaras, M., Frost, S., Haydon, D., Cuceanu, N., Prescott, L., Kamonka, C., Millband, D., Sathar, M. A. & Simmonds, P. (2000). Phylogenetic analysis of GBV C/hepatitis G virus. *Journal of General Virology* **81**, 769–780.

Strauss, E. G. & Strauss, J. H. (1983). Replication strategies of the single stranded RNA viruses of eukaryotes. *Current Topics in Microbiology and Immunology* **105**, 1–98.

Tanaka, T., Kato, N., Cho, M. J., Sugiyama, K. & Shimotohno, K. (1996). Structure of the 3′ terminus of the hepatitis C virus genome. *Journal of Virology* **70**, 3307–3312.

Tanaka, Y., Mizokami, M., Orito, E., Ohba, K., Kato, T., Kondo, Y., Mboudjeka, I., Zekeng, L., Kaptue, L., Bikandou, B., Mpele, P., Takehisa, J., Hayami, M., Suzuki, Y. & Gojobori, T. (1998). African origin of GB virus C hepatitis G virus. *FEBS Letters* **423**, 143–148.

Xiang, J., Wunschmann, S., Schmidt, W., Shao, J. & Stapleton, J. T. (2000). Full-length GB virus C (hepatitis G virus) RNA transcripts are infectious in primary CD4-positive T cells. *Journal of Virology* **74**, 9125–9133.

Yu, H. Y., Grassmann, C. W. & Behrens, S. E. (1999). Sequence and structural elements at the 3′ terminus of bovine viral diarrhea virus genomic RNA: functional role during RNA replication. *Journal of Virology* **73**, 3638–3648.

Received 17 August 2000; Accepted 10 January 2001